Карен Хао
«Империя ИИ: мечты и кошмары OpenAI Сэма Альтмана»

ПРОЛОГ

БЕГ ЗА ТРОНОМ

В пятницу, 17 ноября 2023 года, около полудня по тихоокеанскому времени Сэм Альтман — генеральный директор OpenAI, золотой мальчик Кремниевой долины и живое воплощение революции генеративного искусственного интеллекта — подключился к Google Meet. На экране его уже ждали четверо из пяти членов совета директоров.

Илья Суцкевер, главный научный сотрудник OpenAI, не стал тянуть. Альтман уволен. Официальное заявление появится через несколько минут.

В тот момент Альтман находился в номере роскошного отеля в Лас-Вегасе. Он приехал на первый за много десятилетий этап «Формулы-1» в городе — блестящее светское событие, куда съехались знаменитости от Рианны до Дэвида Бекхэма. Эта поездка была короткой передышкой в изнурительном графике путешествий, который Альтман поддерживал почти без остановки с тех пор, как примерно год назад OpenAI выпустила ChatGPT. Несколько мгновений он не мог произнести ни слова. Он отвёл взгляд, пытаясь прийти в себя. Затем, пока разговор продолжался, сделал то, что умел делать особенно хорошо: попытался сгладить ситуацию.

— Чем я могу помочь? — спросил он. Совет директоров попросил его поддержать временного генерального директора, которого уже выбрали ему на смену, — Миру Мурати, до этого занимавшую должность технического директора OpenAI. Альтман всё ещё не вполне понимал, что происходит. Всё казалось дурным сном.

Но он согласился. Через несколько минут Суцкевер отправил новую ссылку на Google Meet Грегу Брокману — президенту OpenAI, близкому союзнику Альтмана и единственному члену совета, которого не пригласили на предыдущую встречу. Суцкевер сообщил Брокману, что тот лишается места в совете директоров, хотя свою должность в компании пока сохранит.

Вскоре появилось официальное заявление: «Уход господина Альтмана последовал за тщательным рассмотрением ситуации советом директоров, который пришёл к выводу, что он не всегда был откровенен в общении с советом, что мешало совету надлежащим образом выполнять свои обязанности. Совет больше не уверен в его способности продолжать руководить OpenAI». Со стороны всё выглядело почти абсурдно.

OpenAI находилась на вершине могущества. После запуска ChatGPT в ноябре 2022 года компания превратилась в самую громкую историю успеха Кремниевой долины. ChatGPT стал самым быстрорастущим потребительским приложением в истории. Стоимость стартапа взлетала с такой скоростью, что инвесторы буквально облизывались, а лучшие специалисты индустрии мечтали попасть на борт этой ракеты. Всего за несколько недель до увольнения Альтмана OpenAI оценивали уже почти в **90 миллиардов долларов**. Компания как раз завершала сделку, которая позволяла сотрудникам продать часть своих акций жаждущим войти в капитал инвесторам. А за несколько дней до кризиса OpenAI провела долгожданное и восторженно встреченное мероприятие, представив самую амбициозную линейку продуктов в своей истории.

И для публики человеком, который всё это совершил, был именно Сэм Альтман. Весной и летом он объездил весь мир и превратился едва ли не в поп-звезду: пресса уже сравнивала уровень его известности с Тейлор Свифт. Людей покоряло странное сочетание его неброской внешности, громких заявлений и почти обезоруживающей искренности.

Незадолго до поездки в Лас-Вегас он снова путешествовал по миру и выступал на саммите руководителей APEC, с привычной лёгкостью произнося фразы, мгновенно становившиеся цитатами.

— Почему вы решили посвятить этому свою жизнь? — спросила его Лорен Пауэлл Джобс, основатель и президент Emerson Collective, вдова Стива Джобса. Альтман ответил:

— Думаю, это будет самая преобразующая и полезная технология из всех, что человечество когда-либо создавало. Уже четыре раза за историю OpenAI — последний раз буквально пару недель назад

— мне доводилось находиться в комнате в тот момент, когда мы словно отодвигали завесу невежества и раздвигали границу познания. Возможность присутствовать при этом — величайшая профессиональная честь моей жизни.

Сотрудники OpenAI узнали об увольнении Альтмана практически одновременно со всем остальным миром. Ссылка на заявление молниеносно перелетала с одного телефона на другой. Но сильнее всего людей потрясал разрыв между этой новостью и репутацией Альтмана. К тому времени в OpenAI работало уже около восьмисот человек, и у многих сотрудников оставалось всё меньше возможностей лично общаться с генеральным директором. Однако тот обаятельный Альтман, которого весь мир видел на сценах международных форумов, мало отличался от человека, которого сотрудники встречали на общих собраниях, корпоративных мероприятиях или — когда он не путешествовал — просто в офисе. Машина слухов заработала мгновенно.

Сотрудники судорожно листали X — бывший Twitter, — выискивая хоть малейшую зацепку. И вдруг кто-то в офисе ухватился за объяснение, которое на секунду показалось самым логичным, и крикнул: — **Альтман баллотируется в президенты!** На несколько мгновений напряжение спало. Потом все поняли, что нет. И догадки вспыхнули с новой силой — теперь уже вперемешку со страхом. Что сделал Альтман? Что-то незаконное? Может быть, дело связано с его сестрой? За месяц до этого она опубликовала ставшие вирусными сообщения, в которых обвиняла брата в насилии. А если не преступление — то какой-нибудь серьёзный этический проступок? Может быть, дело в других инвестициях Альтмана? Или в его переговорах с саудовскими инвесторами о финансировании нового проекта по производству чипов для искусственного интеллекта?

Суцкевер тем временем опубликовал сообщение во внутреннем Slack OpenAI. Через два часа состоится виртуальное общее собрание, на котором он ответит на вопросы сотрудников. Один из них впоследствии вспоминал: — **Это были самые долгие два часа в моей жизни.**

Когда собрание началось, на экране рядом появились Суцкевер, Мурати и оставшиеся руководители OpenAI. Они выглядели скованно. Неподготовленно. Трансляцию смотрели сотрудники и в офисе, и из дома.

Суцкевер был мрачен. В OpenAI его знали как глубокого мыслителя и человека с почти мистическим складом ума. Он постоянно говорил о технологиях в духовных категориях и делал это с такой неподдельной серьёзностью, что одних это трогало, а других раздражало. При этом в нём была и совершенно другая сторона. Он мог вести себя как большой ребёнок — добродушный, немного чудаковатый. Приходил в офис в футболках с животными. Любил рисовать их сам: умильного кота, умильных альпак, умильного огнедышащего дракона. Рядом — абстрактные лица, обыденные предметы. Некоторые из его любительских картин висели прямо в офисе OpenAI. На одной были изображены три цветка, распустившиеся в форме логотипа компании — символ того, что Суцкевер постоянно призывал сотрудников создавать: **«Множество любящих человечество AGI».** Теперь перед ним стояла совсем иная задача.

Нужно было успокоить сотни встревоженных сотрудников, которые с бешеной скоростью присылали вопросы через общий онлайн-документ. Но Суцкевер был не лучшим человеком для такой роли. Он умел мыслить. Умел говорить о будущем человечества. А вот доносить неприятные новости так, чтобы люди действительно их приняли, — значительно хуже.

Мурати зачитала первый вопрос: — Был ли какой-то конкретный инцидент, который привёл к этому решению? Суцкевер ответил:

— Многие вопросы в документе касаются подробностей: что именно произошло, когда, как, кто конкретно… Хотел бы я рассказать детали. Но не могу. Если кому-то хочется понять больше, добавил он, следует внимательно прочитать официальное заявление. — Там на самом деле очень многое сказано. Возможно, стоит прочесть его несколько раз.

Сотрудники были ошеломлены. Только что им сообщили новость, способную перевернуть всю компанию. Разве они — люди, которых происходящее затронет непосредственно, — не заслуживали знать больше, чем случайный читатель пресс-релиза? Мурати продолжила.

Что теперь будет с отношениями с Microsoft? Это был вопрос не праздный. Microsoft была крупнейшим инвестором OpenAI и обладала эксклюзивной лицензией на её технологии. Кроме того, именно Microsoft предоставляла компании всю вычислительную инфраструктуру. Без неё практически всё, чем занималась OpenAI, остановилось бы: исследования, обучение моделей искусственного интеллекта, запуск новых продуктов. Мурати сказала, что никаких серьёзных последствий не ожидает. Они только что разговаривали с генеральным директором Microsoft Сатьей Наделлой и техническим директором Кевином Скоттом. — Они полностью привержены нашей работе, — заверила она.

Следующий болезненный вопрос: А что будет со сделкой по продаже сотрудниками своих акций? Для людей, проработавших в OpenAI определённый срок, она могла означать **миллионы долларов**. Закрытие сделки ожидалось совсем скоро. Некоторые сотрудники уже запланировали покупку недвижимости. Кто-то даже успел её приобрести. Главный операционный директор Брэд Лайткэп замялся: — Что касается сделки… ну… посмотрим.

Он добавил, что находится на связи с инвесторами, которые её возглавляют, а также с крупнейшими нынешними акционерами. — Все они заверили нас в своей неизменной поддержке компании. Но после ещё нескольких расплывчатых ответов кто-то снова попытался задать главный вопрос. **Что всё-таки сделал Сэм?** Это связано с его работой? Или с личной жизнью? Суцкевер снова сослался на пресс-релиз. — Ответ там, — сказал он.

Мурати прочитала следующий вопрос: — На вопросы о подробностях когда-нибудь ответят? Или никогда?

Суцкевер ответил: — **Не ждите слишком многого.**

По мере того как собрание продолжалось, а ответы Суцкевера всё сильнее расходились с настроением людей, тревога сотрудников стремительно превращалась в гнев.

Суцкевер попытался придать происходящему почти философский смысл: — Когда группа людей вместе проходит через тяжёлое испытание, они часто становятся сплочённее и ближе друг к другу. Это трудное время сделает нас ещё более единой командой — а значит, и более продуктивной.

В документе появился новый вопрос: — Как можно говорить о том, что кризис должен нас объединить, одновременно отказываясь от прозрачности? Обычно правда — необходимое условие примирения.

Суцкевер признал: — Ну… справедливое замечание. Ситуация не идеальна. Мурати попыталась погасить нарастающее раздражение: — Наша миссия гораздо больше любого из нас.

Лайткэп поддержал её: партнёры, клиенты и инвесторы OpenAI один за другим подчёркивали, что по-прежнему верят в миссию компании. — Теперь, пожалуй, наша обязанность следовать этой миссии стала ещё больше.

Суцкевер снова попытался вдохновить людей: — У нас есть всё необходимое. Абсолютно всё. Вычислительные мощности. Исследования. Наши прорывы потрясают воображение. Когда вам тревожно, когда вам страшно — вспоминайте об этом. Представьте мысленно размеры нашего вычислительного кластера. Просто вообразите все эти GPU, работающие вместе.

Появился новый вопрос: — Не опасаемся ли мы враждебного захвата компании посредством давления на нынешних членов совета директоров? Мурати прочитала его вслух. В голосе Суцкевера впервые прозвучало раздражение:

— Враждебного захвата? Он сделал ударение на словах. — Некоммерческий совет директоров OpenAI действовал исключительно в соответствии со своей задачей. Это не враждебный захват. Ни в коем случае. Я совершенно не согласен с постановкой этого вопроса.

В тот вечер несколько сотрудников собрались дома у коллеги на вечеринку, запланированную ещё до увольнения Альтмана. Пришли и сотрудники других компаний в области искусственного интеллекта — в том числе Google DeepMind и Anthropic.

Незадолго до начала мероприятия всем гостям пришло сообщение: **«Сегодня у нас появится ещё одна тематическая комната: “Комната, где не говорят об OpenAI”. До встречи!»** В итоге почти никто там не задерживался. Все хотели говорить именно об OpenAI. Тем же днём Грег Брокман объявил, что в знак протеста увольняется.

Сатья Наделла был в ярости: Microsoft сообщили об увольнении Альтмана буквально за несколько минут до официального объявления. Тем не менее его публичное заявление было выверено до последнего слова. Он написал, что Microsoft имеет долгосрочное соглашение с OpenAI, располагает полным доступом ко всему необходимому для реализации своей инновационной стратегии и по-прежнему привержена партнёрству с компанией, Мирой Мурати и её командой. Но слухи продолжали множиться.

А затем пришла новая новость. Из OpenAI ушли ещё трое ведущих исследователей: Якуб Пахоцкий и Шимон Сидор — одни из самых давних сотрудников компании, — а также профессор MIT Александр Мадры, присоединившийся к OpenAI сравнительно недавно. Для многих это стало тревожным сигналом. Компания теряла руководство. Теряла талантливых специалистов. Следующими могли испугаться инвесторы. Сделка по продаже акций могла сорваться. А в худшем случае могла развалиться сама OpenAI. На вечеринке настроение становилось всё мрачнее. Если сделка рухнет, сотрудники потеряют огромную финансовую отдачу за годы напряжённого труда. Но ещё страшнее было представить исчезновение самой компании — вместе со всеми вложенными в неё усилиями и всеми надеждами.

Тем временем той же ночью совет директоров и оставшееся руководство OpenAI проводили одно совещание за другим. И с каждым новым разговором атмосфера становилась всё более враждебной. Показное единство, которое Суцкевер и остальные руководители пытались продемонстрировать во время общего собрания, развалилось почти сразу после окончания трансляции. Многие из тех, кто сидел рядом с ним на экране, на самом деле были ошеломлены не меньше рядовых сотрудников: об увольнении Альтмана они узнали лишь за несколько минут до официального сообщения. А после неудачного выступления Суцкевера их раздражение переросло в ярость. Они потребовали встречи с остальными членами совета директоров. Около дюжины руководителей, включая Мурати и Лайткэпа, собрались в конференц-зале офиса.

Суцкевер подключился дистанционно вместе с тремя независимыми директорами: Адамом Д'Анджело — сооснователем и генеральным директором сайта вопросов и ответов Quora; Ташей Макколи — предпринимательницей и старшим внештатным научным сотрудником RAND; и Хелен Тонер — исследовательницей австралийского происхождения из Центра безопасности и новых технологий Джорджтаунского университета, CSET.

На них обрушился шквал вопросов. Но четверо членов совета снова и снова отказывались раскрывать подробности, ссылаясь на юридические обязательства по сохранению конфиденциальности. Некоторые руководители уже не скрывали ярости. — Вы утверждаете, что Сэму нельзя доверять, — резко сказала вице-президент по глобальным вопросам Анна Маканджу, которая неоднократно сопровождала Альтмана во время его всемирного турне. — Но это совершенно не соответствует нашему опыту работы с ним.

Руководители потребовали, чтобы члены совета немедленно ушли в отставку и передали свои места трём сотрудникам компании. В противном случае они угрожали уволиться все вместе. Главный стратегический директор Джейсон Квон, юрист и бывший генеральный юрисконсульт OpenAI, пошёл ещё дальше. По его словам, отказ совета уйти в отставку фактически незаконен: если из-за этого компания развалится, члены совета нарушат свои фидуциарные обязанности. Совет директоров отверг этот довод.

Его члены утверждали, что перед увольнением Альтмана тщательно консультировались с юристами и действовали строго в рамках возложенных на них полномочий. В конце концов, OpenAI была **не обычной компанией**. И её совет директоров был **не обычным советом директоров**. Структуру организации во многом разработал сам Альтман. Она предоставляла совету чрезвычайно широкие полномочия и обязывала его действовать не в интересах акционеров OpenAI, а в интересах **миссии компании**: обеспечить, чтобы AGI — искусственный общий интеллект — служил на благо человечества. Сам Альтман много лет называл право совета директоров уволить его важнейшим механизмом управления OpenAI. Хелен Тонер напомнила об этом предельно жёстко: — **Даже если это решение уничтожит компанию, оно всё равно может соответствовать нашей миссии.**

Руководители тут же передали её слова сотрудникам. И в пересказе они зазвучали ещё страшнее: **Тонер всё равно, если компания будет уничтожена.** Некоторые начали подозревать: а может быть, именно этого она и добивается? На вечеринке один человек, представив, что потеряет все свои акции, не выдержал и заплакал.

На следующий день, в субботу 18 ноября, десятки людей, включая сотрудников OpenAI, собрались в особняке Альтмана стоимостью **27 миллионов долларов** и ждали новостей.

Трое ушедших старших исследователей — Пахоцкий, Сидор и Мадры — встретились с Альтманом и Брокманом, чтобы обсудить возможность создать новую компанию и продолжить работу уже там. Одних сотрудников эти слухи напугали ещё сильнее. Появление нового конкурента OpenAI могло окончательно дестабилизировать компанию. Другие, наоборот, увидели надежду. Если Альтман действительно создаст новую компанию, **они уйдут за ним**.

Оставшееся руководство OpenAI предъявило совету директоров ультиматум. Срок — **17:00 субботы по тихоокеанскому времени**. Верните Альтмана. Уходите в отставку. Или компания столкнётся с массовым исходом сотрудников. Совет отказался. Все выходные его члены лихорадочно обзванивали знакомых — порой посреди ночи, — пытаясь найти хоть кого-нибудь, кто снимет трубку и сможет помочь. Гнев сотрудников и инвесторов нарастал. Мурати больше не хотела оставаться временным генеральным директором. Теперь совету требовалось либо найти ей замену — человека, который сумеет вернуть компании хоть какую-то стабильность, — либо подобрать новых членов совета, способных противостоять Альтману, если тот всё-таки вернётся.

Вечером, когда ультиматум уже истёк, Джейсон Квон разослал сотрудникам служебную записку: — Мы по-прежнему работаем над разрешением ситуации и сохраняем оптимизм. И уточнил: — Под разрешением ситуации мы понимаем возвращение Сэма, Грега, Якуба, Шимона и Александра.

Вскоре Альтман написал в X в своём фирменном стиле — строчными буквами: **«я так сильно люблю команду openai».**

Десятки сотрудников начали перепубликовывать его сообщение, добавляя сердечко.

В воскресенье Альтман и Брокман вернулись в офис OpenAI — вести переговоры о своём возвращении. В течение дня туда стекалось всё больше сотрудников. Они сидели и ждали. Большинство работников, руководителей и членов совета директоров не спали нормально уже больше тридцати шести часов. Всё начинало сливаться в один бесконечный, мутный день. Альтман опубликовал селфи: сжатые губы, нахмуренные брови, в руке — гостевой пропуск. Подпись: **«первый и последний раз, когда я ношу такую штуку».** Руководство выставило совету новый срок — снова до пяти вечера. Вернуть Альтмана.

И уйти в отставку. Теперь давление шло со всех сторон. Microsoft, инвесторы OpenAI и влиятельнейшие фигуры Кремниевой долины публично встали на сторону Альтмана. В прессу утёк сценарий дальнейших действий. Если совет не отменит своё решение, сотрудники уйдут массово. Microsoft перекроет доступ к вычислительной инфраструктуре. Инвесторы подадут иски. Всё это вместе сделает существование OpenAI без Альтмана практически невозможным.

Но совет продолжал сопротивляться. К девяти вечера — спустя несколько часов после очередного истёкшего ультиматума — Суцкевер опубликовал от имени совета длинное сообщение в Slack. **Альтман не вернётся.** Новым временным генеральным директором OpenAI назначен Эмметт Шир, бывший глава Twitch. Через пять минут Шир и Суцкевер должны были приехать в офис и рассказать сотрудникам о новом видении будущего компании. Суцкевер написал:

«Совет твёрдо убеждён, что принятое решение — единственный путь, позволяющий продвигать и защищать миссию OpenAI. Проще говоря, поведение Сэма и недостаток прозрачности в его взаимодействии с советом подорвали способность совета эффективно осуществлять надзор за компанией так, как того требуют возложенные на него полномочия». Slack взорвался.

Ответы сотрудников появлялись один за другим: **«И с какой, блядь, армией вы это сделаете?» «Вы бредите». «Эмметт будет генеральным директором ничего».** Около двухсот сотрудников демонстративно покинули офис, бойкотируя выступление. Мурати поспешно увела руководителей из здания. Когда Шир и Суцкевер прибыли, в зале осталось лишь около дюжины человек. К Суцкеверу подошла Анна Брокман, жена Грега.

Четыре года назад именно Суцкевер проводил гражданскую церемонию их бракосочетания. Теперь Анна плакала. Она обняла его и стала умолять изменить решение. Многие ушедшие из офиса сотрудники разъехались по домам коллег, чтобы вместе переждать эту ночь.

Сотни человек вступили в группу Signal, где появлялись последние новости. Поздно вечером Сатья Наделла объявил: **Microsoft нанимает Сэма Альтмана и Грега Брокмана руководить новым подразделением искусственного интеллекта.** Вслед за этим распространилась ещё более важная информация: **каждому сотруднику OpenAI, который захочет уйти вслед за Альтманом, гарантировано место в Microsoft.**

Настроение изменилось мгновенно. Страх сменился вызовом. Теперь у сотрудников был запасной аэродром — а значит, появились и новые рычаги давления на совет директоров и Шира. В доме одного из сотрудников собрались больше сотни работников OpenAI. Руководители и ведущие исследователи начали составлять открытое письмо, чтобы ещё сильнее надавить на совет. Ультиматум был сформулирован предельно ясно: если Альтман не будет восстановлен в должности, а совет директоров не уйдёт в отставку, они **все немедленно уволятся и перейдут в Microsoft**.

Письмо распространяли по частным каналам. Звонили сотрудникам, которых не было рядом, и просили подписать. Когда число подписей достигло критической массы, к письму начали присоединяться остальные — теперь уже опасаясь, что отсутствие их имени само по себе вызовет вопросы. Менее чем за сутки документ подписали **более 700 из примерно 770 сотрудников OpenAI**. Десятки людей один за другим отправляли совету идентичные электронные письма. Сотни сотрудников публиковали в X одну и ту же фразу: **«OpenAI — ничто без своих людей».**

А затем, глубокой ночью, произошло нечто неожиданное. В списке подписавших появилось имя: **Илья Суцкевер.**

Вскоре он выступил публично: «Я глубоко сожалею о своём участии в действиях совета директоров. Я никогда не хотел причинить вред OpenAI. Я люблю всё, что мы создали вместе, и сделаю всё возможное, чтобы снова объединить компанию». Во вторник, 21 ноября, руководство связывалось с советом директоров из дома Альтмана. Шёл пятый день кризиса.

Недосып и изнеможение ломали людей. Приближался День благодарения. И наконец отчаяние с обеих сторон открыло дорогу к компромиссу. В течение дня обсуждались самые разные варианты. Альтман и Брокман, прежде настаивавшие на возвращении в OpenAI вместе с местами в совете директоров, наконец согласились отказаться от кресел.

Совет, в свою очередь, признал очевидное: **без Альтмана сохранить компанию уже невозможно.**

И согласился на его возвращение. Поздно вечером стороны договорились о составе трёх независимых директоров. Адам Д'Анджело оставался. Хелен Тонер и Таша Макколи уходили.

Их места занимали Брет Тейлор — бывший сопредседатель Salesforce и бывший технический директор Facebook — и Ларри Саммерс, бывший министр финансов США и бывший президент Гарвардского университета. Позже новый совет должен был пополниться другими членами. Альтман соглашался пройти внутреннее расследование. Но сейчас важнее всего было другое: **продемонстрировать единство.** Стабильность. Примирение.

Через два дня Альтман опубликует нарочито миролюбивое сообщение: **«только что провёл несколько замечательных часов с @adamdangelo. счастливого дня благодарения от наших семей вашим».** Ещё через десять дней Брокман выложит фотографию: он и Суцкевер стоят, обняв друг друга за плечи и широко улыбаясь.

А в офисе штатный художник OpenAI подключит генератор изображений DALL-E к цветному принтеру и начнёт печатать крошечные калейдоскопические наклейки в форме сердца. Рядом появится огромное розовое сердце с надписью: **«OpenAI — ничто без своих людей».** В декабре Альтман скажет Тревору Ноа в подкасте, что эта неделя была **вторым худшим переживанием в его жизни**. Хуже была только смерть отца.

А в январе сделка по продаже акций сотрудников всё-таки завершится. OpenAI будет оценена в **86 миллиардов долларов**. Но всё это случится потом. А вечером вторника, 21 ноября, люди просто

праздновали. Как только объявили о возвращении Альтмана и новом соглашении, сотрудники хлынули обратно в офис.

Обнимались. Плакали. Врубили музыку. Кто-то притащил дым-машину. Она сработала так хорошо, что включилась пожарная сигнализация. Никого это не остановило. Праздник продолжался. Брокман поднял телефон и сделал групповое селфи. На снимке — толпа людей в состоянии почти безумной эйфории: измученные, невыспавшиеся, счастливые оттого, что пережили катастрофу и каким-то образом остались целы.

Он выложил фотографию с двумя словами: **«мы вернулись».** Новость об изгнании Альтмана застала меня прямо посреди интервью для этой книги. Я отключила звук на телефоне и ещё не подозревала, в какую безумную неделю вот-вот окажусь втянута.

Минут через двадцать я взглянула на экран, чтобы посмотреть время, — и увидела целую россыпь пропущенных уведомлений. Так начались несколько туманных, наполненных адреналином дней, когда я отчаянно пыталась понять, что же происходит. В последующие недели друзья, родственники и журналисты снова и снова задавали мне один и тот же вопрос: **Что всё это означает?** Если вообще что-нибудь означает. Была ли эта череда увольнений, ультиматумов и возвращений всего лишь увлекательной кремниеводолинской драмой? Или последствия почувствуем и мы?

К тому времени я следила за OpenAI уже пять лет. В 2019 году я стала первым журналистом, получившим широкий доступ к компании и написавшим о ней большой профиль. Поэтому я воспринимала произошедшее не как очередную мелочную борьбу за власть между богатыми людьми Кремниевой долины. Эта драма высветила один из самых насущных вопросов нашего поколения: **кто и как должен управлять искусственным интеллектом?** ИИ — одна из наиболее значимых технологий нашей эпохи. Чуть больше чем за десятилетие он перестроил саму основу интернета и превратился в вездесущего посредника почти всех наших действий в цифровом мире. А теперь, за ещё более короткий срок, он готовится перекроить множество важнейших общественных институтов: здравоохранение, образование, право, финансы, журналистику, государственное управление. Будущее искусственного интеллекта — то, какой именно станет эта технология, — неразрывно связано с **нашим собственным будущим**. Поэтому вопрос о том, кто будет управлять ИИ, на самом деле звучит иначе:

как сделать так, чтобы будущее стало лучше, а не хуже? С самого начала OpenAI представляла себя как смелый эксперимент, призванный дать ответ именно на этот вопрос. Компанию основала группа людей, среди которых были Илон Маск и Сэм Альтман, при финансовой поддержке других миллиардеров, включая Питера Тиля. Она должна была стать чем-то большим, чем исследовательская лаборатория. И большим, чем обычная компания. Основатели объявили о радикальной цели: создать так называемый **искусственный общий интеллект — AGI**, то есть, по их представлению, самую могущественную форму искусственного интеллекта из всех когда-либо существовавших. Но создать его не ради прибыли акционеров. **Ради всего человечества.** Именно поэтому Маск и Альтман оформили OpenAI как некоммерческую организацию и пообещали выделить на её работу миллиард долларов. Компания не должна была заниматься коммерческими продуктами. Только исследованиями. Только благими намерениями.

Только созданием такого AGI, который откроет человечеству путь к глобальной утопии — а не к её противоположности. Маск и Альтман также обещали максимально открыто публиковать результаты исследований и широко сотрудничать с другими организациями. Если ваша цель — благо всего мира, рассуждали они, то открытость — отсюда и название **OpenAI** — и демократическое участие в развитии технологии имеют принципиальное значение. Через несколько лет руководство пошло ещё дальше. Оно даже публично пообещало в случае необходимости **пожертвовать собственным первенством**. Руководители OpenAI писали, что опасаются превращения заключительного этапа разработки AGI в гонку, в которой ради скорости участники будут пренебрегать безопасностью. Поэтому, если другой проект приблизится к созданию безопасного и полезного AGI раньше OpenAI, компания обещала прекратить с ним конкурировать и начать ему помогать. Звучало почти беспрецедентно. Но к тому моменту, когда я впервые начала подробно изучать OpenAI, приверженность этим идеалам уже стремительно исчезала. Всего через полтора года

после создания организации её руководство осознало: выбранный ими путь развития искусственного интеллекта потребует **невероятных денег**. До этого Маск и Альтман в качестве сопредседателей держались относительно в стороне от повседневного управления. Теперь оба захотели стать генеральным директором. Победил Альтман. В начале 2018 года Маск покинул OpenAI. И забрал свои деньги. Оглядываясь назад, трудно не увидеть в этом расколе первый серьёзный признак того, что OpenAI была не столько альтруистическим проектом, сколько **проектом человеческих амбиций и эго**.

Потеря главного спонсора поставила OpenAI в тяжёлое финансовое положение. Чтобы найти выход, Альтман перестроил юридическую архитектуру организации. Внутри некоммерческой структуры он создал коммерческое подразделение — **OpenAI LP**. Теперь оно могло привлекать капитал. Выпускать коммерческие продукты. И приносить инвесторам прибыль — почти как любая обычная компания. Через четыре месяца, в июле 2019 года, OpenAI объявила о появлении нового инвестора на миллиард долларов. Им стала Microsoft.

Вскоре после этого, в августе 2019 года, я впервые пришла в офис OpenAI. Три дня я провела внутри компании, разговаривая с сотрудниками. Провела десятки интервью. И увидела, как эксперимент с идеалистической системой управления начинает распадаться. OpenAI становилась всё более конкурентной. Всё более закрытой. Всё более замкнутой на себе. Она словно начинала бояться внешнего мира — и одновременно пьянела от осознания того, что контролирует технологию потенциально колоссальной силы.

Идеи прозрачности и демократии исчезали. Как и разговоры о самопожертвовании и сотрудничестве. У руководителей OpenAI постепенно осталась одна навязчивая цель: **первыми создать искусственный общий интеллект.** И создать его **по собственному образу и подобию**. В последующие четыре года OpenAI превратилась почти во всё то, чем когда-то обещала **не становиться**. Некоммерческой организация оставалась преимущественно по названию. Она агрессивно коммерциализировала такие продукты, как ChatGPT, и добивалась прежде немыслимых оценок стоимости.

Становилась всё более секретной. Она не только закрывала доступ к собственным исследованиям, но и меняла правила всей отрасли, выводя значительную часть разработок в области ИИ из-под общественного контроля. Она сама запустила именно ту **гонку ко дну**, против которой когда-то предостерегала. Коммерциализация и внедрение искусственного интеллекта ускорялись с невероятной скоростью — прежде чем удалось устранить его опасные недостатки и понять, как эта технология способна усиливать и эксплуатировать уже существующие расколы общества. Тем временем конфликты между руководителями и сотрудниками становились всё ожесточённее. Различные группировки внутри компании боролись за контроль, каждая пытаясь перестроить OpenAI в соответствии со своим представлением о будущем.

Увольнение Альтмана в ноябре 2023 года и его почти мгновенное возвращение стали окончательным доказательством: **эксперимент с управлением провалился.** И дело было не только в том, что некоммерческий совет директоров отступил под давлением денег, уничтожив последние остатки альтруистического фасада организации. Произошедшее продемонстрировало кое-что гораздо важнее. Будущее искусственного интеллекта сегодня формируется в ходе борьбы за власть между **ничтожным числом представителей элиты Кремниевой долины**. Даже если бы события развернулись иначе и совету удалось заменить Альтмана, главная проблема осталась бы прежней: решение колоссальной важности принималось **за закрытыми дверями**. За пределами крошечной группы сверхбогатых технооптимистов, их самых яростных идеологических противников и многомиллиардного технологического гиганта даже собственные сотрудники OpenAI почти ничего не знали о том, какая судьба им уготована.

Я начала писать об искусственном интеллекте задолго до того, как OpenAI и ChatGPT стали почти синонимами самого понятия ИИ. Я наблюдала, как технология развивается через беспорядочный, полный проб и ошибок процесс науки и инноваций. Как исследователи проверяют новые идеи. Как представляют самые впечатляющие результаты на переполненных конференциях. Как затем внедряют

эти разработки в коммерческие продукты крупнейших компаний мира — Google, Facebook, Alibaba, Baidu. Я прочла сотни научных статей и беседовала с учёными, инженерами и руководителями, пытаясь понять их мировоззрение и решения — и увидеть, какой отпечаток эти взгляды оставляют на конструкции и способах применения технологий.* По мере того как искусственный интеллект распространялся по миру, я следила и за едва заметными, и за драматическими переменами, которые он приносил в жизнь людей и целых сообществ. Я побывала на пяти континентах. В Колумбии и Кении встречалась с людьми, которые из-за экономического кризиса пошли работать на индустрию ИИ, размечая данные, — и обнаружили себя в условиях, пугающе напоминавших долговое рабство. В Аризоне и Чили я разговаривала с местными политиками и активистами, встревоженными появлением целых теневых мегаполисов дата-центров, пожирающих драгоценные запасы воды, от которых зависят их дома и города. За годы работы я поняла две вещи. Во-первых, **искусственный интеллект — это не одна технология**.

Это множество различных технологий, которые постоянно меняют форму и развиваются — причём направление этого развития определяется не только техническими достоинствами. На него влияют идеология людей, создающих технологии. Мода. Ажиотаж. Коммерческие интересы. Сегодня всё внимание досталось ChatGPT, большим языковым моделям и генеративному ИИ. Но это всего лишь **одно из возможных воплощений искусственного интеллекта**. И воплощение это отражает весьма конкретное — и поразительно узкое — представление о том, как устроен мир и каким он должен быть. Не было никакой неизбежности ни в том, что именно такая форма ИИ выйдет на первый план, ни даже в том, что она вообще появится. Она стала результатом **тысяч субъективных решений**, принятых теми людьми, которым было позволено находиться в комнате, где принимаются решения. И будущие поколения технологий ИИ точно так же не предопределены. А значит, мы снова возвращаемся к главному вопросу: **кому будет позволено их формировать?** Второе, что я поняла, значительно тревожнее. Нынешняя форма искусственного интеллекта — и направление, в котором она развивается, — ведут нас по опасному пути.

На поверхности генеративный ИИ восхищает. Он может за секунды помочь придумать идею или написать текст. Может стать собеседником глубокой ночью, когда хочется избавиться от одиночества. Возможно, когда-нибудь он действительно настолько повысит производительность труда, что даст мощный толчок экономике. Но когда-то мы точно так же считали Facebook всего лишь местом, куда выкладывают фотографии из отпуска, где находят давно потерянных школьных друзей и где зарождаются позитивные общественные движения. За гладкой, завораживающей оболочкой генеративного ИИ скрывается нечто гораздо более сложное. Под капотом эти модели — **настоящие чудовища масштаба**. Для их создания поглощаются объёмы данных, человеческого труда, вычислительной мощности и природных ресурсов, которые ещё недавно невозможно было вообразить. По одной из оценок, GPT-4 — преемник модели, лежавшей в основе первого ChatGPT, — более чем в **пятнадцать тысяч раз крупнее GPT-1**, выпущенной всего пятью годами раньше. Человеческая и материальная цена этого взрывного роста ложится на огромные слои общества. Причём прежде всего — на самых уязвимых. На людей, которых я встречала по всему миру: работников и жителей сельских районов богатого Глобального Севера; беднейшие сообщества Глобального Юга. Все они сталкиваются с новыми формами незащищённости. А обещанные блага «технологической революции» почти не просачиваются вниз. **Выгоды генеративного ИИ в основном текут вверх.** За все эти годы я нашла лишь одну метафору, которая, как мне кажется, действительно передаёт сущность нынешних властителей искусственного интеллекта.

Империи. Во времена европейского колониализма империи захватывали и выкачивали ресурсы, которые им не принадлежали. Они эксплуатировали труд покорённых народов, заставляя их добывать, выращивать и перерабатывать эти ресурсы ради обогащения метрополии. Они распространяли расистские и дегуманизирующие представления о собственном превосходстве и современности, чтобы оправдать захват суверенитета, грабёж и подчинение — а иногда даже убедить покорённых принять всё это добровольно. Стремление к могуществу оправдывалось необходимостью конкурировать с другими империями. А на гонке вооружений, как известно, правила заканчиваются. В конечном итоге всё это служило одной цели: укреплять власть империи, расширять её границы,

обеспечивать её дальнейший рост. Проще говоря, империи накапливали невероятные богатства, растягиваясь через пространство и века и навязывая миру колониальный порядок — за который все остальные платили огромную цену. Современные **империи ИИ**, разумеется, не прибегают к той открытой жестокости и насилию, которыми была отмечена эпоха колониализма.

Но и они захватывают и извлекают драгоценные ресурсы, необходимые для воплощения их представления об искусственном интеллекте: труды художников и писателей; данные бесчисленных людей, рассказывающих о своей жизни и наблюдениях в интернете; землю, электричество и воду, необходимые гигантским дата-центрам и суперкомпьютерам. Новые империи точно так же эксплуатируют труд людей по всему миру — тех, кто очищает, сортирует, размечает и подготавливает данные, которые затем превращаются в прибыльные технологии искусственного интеллекта. Они рисуют соблазнительные картины будущего. Говорят о современности. И агрессивно убеждают нас, что обязаны победить конкурирующие империи. Так создаётся оправдание для вторжения в частную жизнь, присвоения чужого труда и массовой автоматизации, способной уничтожить огромные пласты содержательной экономической деятельности.

И сегодня именно OpenAI возглавляет наше стремительное движение к этому новому колониальному порядку. Стремясь к расплывчатому идеалу «прогресса», компания сделала бесконечное наращивание масштаба главным правилом новой эпохи искусственного интеллекта. Теперь технологические гиганты соревнуются в том, кто станет ещё больше. Они тратят настолько астрономические суммы, что вынуждены перекраивать и концентрировать собственные ресурсы. Примерно тогда, когда Microsoft инвестировала **10 миллиардов долларов** в OpenAI, она одновременно уволила десять тысяч сотрудников ради сокращения расходов. Увидев, что OpenAI обгоняет её, Google объединила свои лаборатории искусственного интеллекта в Google DeepMind. Когда Baidu бросилась создавать собственный аналог ChatGPT, сотрудникам, работавшим над использованием ИИ для поиска новых лекарств, пришлось приостановить исследования и отдать свои вычислительные чипы разработчикам чат-бота. Сегодняшняя модель развития искусственного интеллекта перекрывает кислород альтернативным направлениям. Число независимых исследователей, не связанных с технологическими компаниями и не получающих от них финансирование, быстро сокращается. А вместе с ними исчезает и разнообразие идей, которые нельзя немедленно превратить в коммерческую выгоду.

Даже сами компании, когда-то финансировавшие широкие и свободные фундаментальные исследования, теперь всё чаще не могут себе этого позволить: счёт за генеративный ИИ слишком велик. Новое поколение учёных вынуждено приспосабливаться к сложившемуся порядку, чтобы оставаться востребованным на рынке труда. То, что ещё недавно казалось беспрецедентным, стало нормой. Сегодня эти империи богаче, чем когда-либо прежде.

Когда в январе 2025 года я заканчивала эту книгу, OpenAI оценивалась уже в **157 миллиардов долларов**. Её конкурент Anthropic приближался к сделке с оценкой в **60 миллиардов**. После заключения партнёрства с OpenAI рыночная капитализация Microsoft выросла втрое и превысила **3 триллиона долларов**. После появления ChatGPT совокупная рыночная стоимость шести крупнейших технологических гигантов увеличилась примерно на **8 триллионов долларов**. И одновременно возникает всё больше сомнений: а насколько велика **реальная экономическая ценность** генеративного ИИ? В июне 2024 года Goldman Sachs подсчитал, что в течение нескольких лет расходы на развитие этой технологии могут достичь **триллиона долларов**, хотя пока, по формулировке авторов отчёта, предъявить в оправдание этих затрат можно удивительно мало. Месяцем позже Upwork Research Institute опросил 2500 работников по всему миру. **96 процентов руководителей высшего звена** были уверены, что генеративный ИИ повысит производительность. Но среди сотрудников, которые реально пользовались этими инструментами, **77 процентов заявили обратное**: ИИ увеличивал их рабочую нагрузку. Отчасти потому, что требовалось время на проверку созданного им контента. Отчасти потому, что начальство, вооружившись ИИ, начинало требовать от людей ещё большей производительности. В ноябре журналисты Bloomberg Парми Олсон и Кэролин Сильверман, анализируя финансовые результаты генеративного ИИ, сформулировали неприятную возможность предельно ясно: технология, которую объявили революционной, возможно, **так никогда и не**

принесёт обещанного масштабного преобразования экономики — а лишь ещё сильнее сосредоточит богатство наверху. Тем временем остальной мир уже начинает прогибаться под растущей человеческой и материальной ценой новой эпохи. В Кении люди получали нищенскую зарплату за то, что отфильтровывали сцены насилия и язык ненависти из систем OpenAI, включая ChatGPT.

Художников вытесняют модели искусственного интеллекта, которые обучались на их произведениях — без разрешения и без оплаты. Журналистика слабеет, пока генеративные технологии производят всё большие объёмы дезинформации. Прямо у нас на глазах повторяется **очень древняя история**. И это только начало. OpenAI не собирается сбавлять скорость. Компания продолжает стремиться ко всё большим масштабам, располагая ресурсами, с которыми мало кто способен сравниться. А вслед за ней движется вся отрасль. Чтобы приглушить растущую тревогу по поводу того, что генеративный ИИ способен дать человечеству **сегодня**, Альтман всё громче говорит о фантастических благах AGI **завтра**.

В сентябре 2024 года он объявил в своём блоге, что скоро наступит **«Эпоха интеллекта»**, эпоха «массового процветания». А сверхинтеллект, предположил он, может появиться уже через **«несколько тысяч дней»**. Будущее, писал Альтман, будет настолько светлым, что никакое сегодняшнее описание не сможет отдать ему должное. Постепенно невероятные достижения станут обыденностью: мы исправим климат, создадим колонии в космосе, раскроем все законы физики. Но к этому моменту AGI во многом превратился в **риторический образ** — универсальное фантастическое оправдание, позволяющее OpenAI стремиться ко всё большему богатству и могуществу. Почти ни у кого другого нет сопоставимого капитала, чтобы финансировать альтернативные пути. OpenAI и несколько её крупнейших конкурентов получают фактическую олигополию на технологию, которую они же продают нам как **ключ к будущему**. И всякий — компания или государство, — кто пожелает получить кусочек этого будущего, будет вынужден обратиться к империям, которые контролируют к нему доступ. Но есть другой путь. **Искусственный интеллект не обязан быть таким, каким он стал сегодня.** Нам вовсе не обязательно принимать идею, будто прогресс требует бесконечного роста масштабов и потребления ресурсов. Очень многое из того, что действительно необходимо нашему обществу, — лучшее здравоохранение, лучшее образование, чистый воздух, чистая вода, быстрый отказ от ископаемого топлива — может быть достигнуто при помощи значительно меньших моделей искусственного интеллекта и самых разнообразных альтернативных подходов. А одного ИИ в любом случае будет недостаточно.

Нам понадобится больше общественной сплочённости. Больше международного сотрудничества. Именно тех вещей, которые нынешнее направление развития искусственного интеллекта, напротив, подрывает. Но **империи ИИ не отдадут свою власть добровольно**. Нам придётся самим вернуть себе право определять будущее этой технологии. И сейчас наступил критический момент, когда это ещё возможно. Когда-то старые империи уступили место более инклюзивным формам управления. Так и мы можем вместе определить будущее искусственного интеллекта. Политики могут принять строгие правила защиты персональных данных и прозрачности, обновить законодательство об интеллектуальной собственности и вернуть людям контроль над их данными и результатами их труда. Правозащитные организации могут добиваться международных трудовых норм, которые обеспечат разметчикам данных гарантированный минимальный заработок и достойные условия труда, а также укрепят права работников во всех отраслях экономики. Организации, финансирующие науку, могут вернуть разнообразие исследованиям в области искусственного интеллекта и поддержать принципиально новые представления о том, **чем вообще может быть эта технология**. И наконец, каждый из нас способен сопротивляться историям, которые OpenAI и индустрия ИИ рассказывают нам, чтобы спрятать растущую социальную и экологическую цену технологии за туманным обещанием прогресса.

Примечание

* В области исследований искусственного интеллекта существует особая практика публикации научных работ. Исследователи часто сразу размещают статьи в открытом бесплатном репозитории **arXiv** (произносится как *archive* — «архив»), а рецензирование в научной конференции или журнале проходит лишь спустя месяцы или даже годы. Иногда работа вообще не проходит формального рецензирования. Эта практика настолько укоренилась, что многие специалисты оценивают и цитируют статьи исходя прежде всего из их влияния на научное сообщество, а не из факта прохождения рецензирования. В этой книге я буду придерживаться того же принципа. В примечаниях в конце книги указано,

ГЛАВА 1.

БОЖЕСТВЕННОЕ ПРАВО

Все уже собрались. Не было только Илона Маска — он, как обычно, опаздывал.

Стояло лето 2015 года. По приглашению Сэма Альтмана несколько человек собрались на закрытый ужин, чтобы поговорить о будущем искусственного интеллекта — и, ни много ни мало, о будущем человечества.

Маск познакомился с Альтманом, который был на четырнадцать лет моложе него, некоторое время назад и остался под хорошим впечатлением. Альтман уже тогда был человеком с громким именем. Он возглавлял знаменитый кремниеводолинский акселератор стартапов Y Combinator, а ещё раньше, в девятнадцать лет, основал свою первую компанию и быстро приобрёл репутацию блестящего стратега, искусного переговорщика и человека с огромными амбициями — даже по меркам Кремниевой долины, где грандиозные замыслы считались нормой. Маску нравилось в нём несколько вещей. Альтман был умён. Целеустремлён. И, что важнее всего, говорил об искусственном интеллекте почти так же, как сам Маск: развивать такую технологию нужно крайне осторожно, а управление ею нельзя пускать на самотёк.

Позднее Маск утверждал в судебных документах, что Альтман словно отражал его собственные взгляды, повторяя всё, что Маск когда-либо говорил на эту тему, чтобы завоевать его доверие. Сам Альтман неоднократно говорил, что в детстве Маск был одним из его героев. Когда Маск лично провёл его по огромному заводу SpaceX в Хоторне, Калифорния, восхищение только усилилось. Позже Альтман вспоминал: «Больше всего мне запомнилось выражение абсолютной уверенности на его лице, когда он говорил об отправке больших ракет на Марс. Я ушёл оттуда с мыслью: “Ага. Значит, вот как выглядит настоящая убеждённость”».

К тому времени Маск уже несколько лет был всерьёз обеспокоен искусственным интеллектом. В 2012 году он познакомился с Демисом Хассабисом — учёным по складу характера и генеральным директором лондонской лаборатории DeepMind Technologies. Вскоре Хассабис приехал к Маску на завод SpaceX. Они сидели в столовой, а вокруг гремели гигантские детали ракет, которые перевозили и собирали неподалёку. И там Хассабис заговорил о возможности появления искусственного интеллекта, который однажды превзойдёт человеческий разум. Такая система, сказал он, вполне может стать угрозой для человечества. И даже любимый запасной план Маска — если что-то пойдёт совсем плохо, переселиться на Марс — здесь не спасёт. Сверхразум, усмехнулся Хассабис, просто последует за людьми в космос. Самому Маску это показалось совсем не смешным. Он вложил в DeepMind пять миллионов долларов — в том числе для того, чтобы иметь возможность следить за развитием компании.

Год спустя, на праздновании своего дня рождения в винодельческой долине Напа, Маск вступил в жаркий и эмоциональный спор со своим давним другом, сооснователем Google Ларри Пейджем. Предмет спора был прост: будет ли искусственный интеллект, превосходящий человека, угрозой? Пейдж не видел в этом трагедии. Для него такой разум был всего лишь следующим этапом эволюции. Когда Маск возразил, Пейдж обвинил его в «видовом шовинизме» — в том, что тот ставит людей выше нечеловеческих форм разума.

После этого Маск почти без остановки начал говорить об экзистенциальной опасности ИИ. На симпозиуме MIT он назвал искусственный интеллект, возможно, «главнейшей экзистенциальной угрозой» для человечества, а его развитие сравнил с «вызовом демона». Он встречался с издателями в Нью-Йорке, обдумывая собственную книгу о рисках, способных привести к исчезновению человечества, включая угрозу искусственного интеллекта. Позднее на одном из мероприятий Stanford AI Salon к нему после выступления подошла молодая исследовательница Тимнит Гебру. Она спросила: — Почему вы настолько одержимы искусственным интеллектом, если изменение климата гораздо очевиднее представляет угрозу существованию человечества? Маск ответил: — Изменение

климата — это плохо. Но оно не убьёт всех. А искусственный интеллект может уничтожить человечество полностью.

В конце 2013 года Маск узнал, что Google собирается купить DeepMind. Он был уверен: ничем хорошим это не закончится. Публично он предупреждал: если Google однажды даст гипотетическому AGI задачу максимизировать прибыль, система может решить, что лучший способ выполнить эту цель — любой ценой устранить конкурентов. По этому поводу Маск язвительно сказал *The New Yorker*, что решение начать с убийства всех исследователей ИИ, работающих у конкурентов, выглядело бы «не самым приятным свойством характера». На одной вечеринке в Лос-Анджелесе Маск на целый час заперся в гардеробной на втором этаже и разговаривал с Хассабисом по Skype, убеждая его отказаться от сделки с Google. Он сказал: — Будущее искусственного интеллекта не должно находиться под контролем Ларри. Но Маск не знал, что процесс уже зашёл очень далеко. Google отправила группу исследователей ИИ на частном самолёте в офис DeepMind для технической оценки компании. Один из старейших и наиболее влиятельных специалистов Google, Джефф Дин, лично просмотрел часть исходного кода DeepMind и дал сделке добро. В январе 2014 года Google официально подтвердила покупку. По сообщениям, цена составила от 400 до 650 миллионов долларов.

После этого Маск начал устраивать собственные ужины, где обсуждал, как можно противостоять Google. В начале 2015 года он даже встретился с президентом США Бараком Обамой, чтобы рассказать об опасностях ИИ, способах сделать его безопаснее и необходимости государственного регулирования. Примерно тогда же Маск вновь встретился с Хассабисом в SpaceX — на первом заседании совета по этике Google DeepMind. Этот орган Ларри Пейдж и Демис Хассабис создали для надзора за ответственным развитием технологий компании. Маск вышел с заседания с убеждением, что совет — фикция. Его тревога превратилась почти в навязчивую идею:

Хассабиса необходимо остановить. Годами Маск изображал его едва ли не суперзлодеем. В его картине мира OpenAI должна была стать добром, противостоящим злу DeepMind. Летом 2016 года, вскоре после основания OpenAI, несколько её сотрудников встретились с Хассабисом. Вернувшись, они сообщили коллегам примерно следующее: Да, DeepMind действительно намерена захватить мир. Похоже, Маск был прав. На следующий год Маск организовал выездную встречу сотрудников OpenAI на заводе SpaceX и снова разразился тирадой о Хассабисе. До DeepMind тот семь лет руководил собственной студией видеоигр. Маск почти кричал: — Он буквально сделал игру, где злой гений создаёт искусственный интеллект, чтобы захватить мир! И люди, чёрт возьми, этого не понимают! Не понимают! А Ларри? Ларри думает, что контролирует Демиса, но сам слишком занят грёбаным виндсёрфингом и не замечает, как Демис собирает в своих руках всю власть!

Для сотрудников DeepMind эта паранойя Маска со временем стала источником развлечения. Да, Хассабис был чрезвычайно амбициозен. Да, временами очень напорист. Но при этом люди описывали его как спокойного и доброжелательного человека. Один бывший исследователь DeepMind вспоминал: — Создание OpenAI выглядело какой-то полуистерической реакцией на довольно мягкого и спокойного человека. Это казалось немного абсурдным.

Среди книг, которые Маск особенно рекомендовал, была *Superintelligence: Paths, Dangers, Strategies* — «Сверхинтеллект: пути, опасности, стратегии» — оксфордского философа Ника Бострома. Главная мысль книги была тревожной. Если искусственный интеллект станет умнее человека, его может оказаться чрезвычайно трудно контролировать. И последствия могут быть катастрофическими. Бостром приводил ставший знаменитым мысленный эксперимент. Допустим, сверхразумной системе поставили простейшую задачу: производить скрепки. Теоретически она может решить, что люди мешают выполнению этой задачи, потому что потребляют ресурсы, которые иначе можно было бы использовать для производства ещё большего количества скрепок. Отсюда возникал главный вопрос: как сделать так, чтобы сверхразум не исполнял наши инструкции буквально, разрушая всё вокруг? Бостром предлагал идею «выравнивания», или alignment: искусственный интеллект должен быть согласован с человеческими ценностями и способен понимать не только буквальную команду, но и более широкие намерения человека.

Именно из таких идей позднее выросло целое направление исследований безопасности ИИ, которое OpenAI сделает одной из своих главных тем. Своей огромной аудитории в Twitter Маск рекомендовал книгу просто: «Стоит прочитать». Однако в январе 2023 года всплыло старое электронное письмо Бострома середины 1990-х, которое заставило многих иначе взглянуть уже на его собственные ценности. В нём он использовал откровенно расистскую формулировку и утверждал, что считает её истинной. Позднее Бостром извинился, назвал письмо отвратительным и заявил, что оно не отражает его сегодняшних взглядов.

Альтман казался Маску человеком того же склада. Его тоже занимала мысль о катастрофах и о том, как к ним подготовиться. В 2016 году Альтман рассказал журналисту Тэду Френду из *The New Yorker*, что в случае конца света намерен бежать в Новую Зеландию вместе с близким другом и наставником — миллиардером Питером Тилем. В той же статье Тиль охарактеризовал Альтмана как человека «в культурном смысле очень еврейского: оптимиста, но одновременно выживальщика, который всегда помнит, что всё может пойти ужасно плохо». Через два года Альтман скажет Bloomberg, что тогда лишь шутил. Однако аварийный рюкзак у него действительно был. Особенно он опасался новых биологических вирусов. В запасе были: противогазы, антибиотики, вода, батарейки, палатка и оружие. А в феврале 2015 года Альтман написал в своём блоге, что согласен с Маском: сверхразум, вероятно, является величайшей угрозой дальнейшему существованию человечества. Да, разрушительный искусственно созданный вирус, по его мнению, был более вероятным сценарием. Но даже он вряд ли уничтожит всех людей во Вселенной. В скобках Альтман вновь рекомендовал книгу Бострома как лучшее из прочитанного им на эту тему.

Через несколько месяцев, в мае 2015 года, Альтман написал Маску. Он размышлял, возможно ли вообще остановить человечество от создания искусственного интеллекта. Его собственный ответ: почти наверняка нет. А если ИИ всё равно будет создан, рассуждал Альтман, неплохо бы, чтобы первым это сделал кто-нибудь кроме Google. Он предложил создать в Y Combinator нечто вроде: «Манхэттенского проекта для искусственного интеллекта». Но с принципиальным отличием. Технология должна была принадлежать не частной компании, а всему миру — через некоммерческую организацию. Альтман также подчёркивал, что проект будет строго соблюдать и даже активно поддерживать государственное регулирование. Маск ответил коротко:

— Обсудить определённо стоит. В июне Альтман прислал новое письмо с подробностями. Миссия, писал он, должна состоять в создании первого общего искусственного интеллекта и использовании его для расширения возможностей каждого отдельного человека. Самым безопасным будущим ему казалась не централизованная система, а распределённая. И главное: безопасность должна стать фундаментальным требованием с самого начала. Затем Альтман предложил структуру управления. Он и Маск войдут в совет. К ним присоединятся ещё три человека. Технология будет принадлежать фонду и использоваться «на благо мира». А если будет непонятно, что именно означает «благо мира», решение примут эти пять человек. В этом раннем замысле уже присутствовал один примечательный парадокс: технология должна была принадлежать всему человечеству —

но определять, что именно полезно человечеству, собирались пять человек. Альтман попросил Маска по возможности встречаться с командой примерно раз в месяц. Его участие, объяснял Альтман, поможет привлечь лучших специалистов. Если времени у Маска не будет, даже публичной поддержки окажется достаточно — одно его имя значительно облегчит набор людей.

Ответ Маска был лаконичен: — Согласен со всем.

После этого Альтман пригласил Маска на тот самый закрытый ужин о будущем искусственного интеллекта и человечества. Там он хотел познакомить его с группой ведущих инженеров и исследователей, которых надеялся привлечь к новому проекту. Как только Маск подтвердил участие, место ужина изменили на более престижное. Встречу перенесли в любимый Маском Rosewood Hotel — роскошный комплекс площадью шестнадцать акров, где ночь стоила около тысячи долларов. Отель располагался на живописной Sand Hill Road — улице, утопающей в деревьях и буквально окружённой

офисами венчурных фондов. Фактически это была одна из главных артерий Кремниевой долины. Закрытый зал выходил на балкон с видом на бассейн, окружённый итальянскими кипарисами и садовыми розами. Маск появился более чем на час позже назначенного времени. Остальные уже давно его ждали. Среди гостей были: Сэм Альтман,

Грег Брокман, Дарио Амодеи и Илья Суцкевер. Вскоре именно эти люди станут ключевыми руководителями новой некоммерческой организации. Маск придумает ей название, которое должно было выразить сам дух их миссии: OpenAI — «Открытый искусственный интеллект». Ирония истории заключалась в другом.

Со временем почти все люди, сидевшие за тем столом, покинут организацию после конфликтов с Альтманом и его представлением о будущем ИИ. Когда отношения Маска и Альтмана окончательно разрушатся, а Альтман сам станет новой главной звездой Кремниевой долины, изменится и публичная версия его взглядов. В 2023 году он скажет Business Insider: — Сейчас я твёрдо принадлежу к лагерю тех, кто считает, что ИИ будет инструментом. Хотя будущие люди и само человеческое общество станут совсем другими, и у нас есть возможность сознательно подумать о том, как именно спроектировать это будущее.

Маск же постепенно придёт к убеждению: Альтман использовал его, чтобы подняться на вершину.

Впрочем, подобное наблюдение сопровождало Альтмана почти всю жизнь. Его наставник Пол Грэм однажды знаменитым образом сказал: «Можно сбросить Сэма с парашютом на остров каннибалов, вернуться через пять лет — и он уже будет там королём».

Позднее Грэм сформулировал ту же мысль ещё прямее: «Сэм чрезвычайно хорошо умеет становиться влиятельным».

СЭМ АЛЬТМАН

Сэмюэл Харрис Гибстайн Альтман родился 22 апреля 1985 года в Чикаго. Он был первым сыном в еврейской семье. Его мать, Конни Гибстайн, была врачом. Её отец Марвин тоже был врачом — педиатром, который после Второй мировой войны служил в Германии как военный медик армии США. Конни получила сразу два образования — медицинское и юридическое, что для женщины её поколения было весьма необычно. В итоге она выбрала дерматологию: стабильный заработок и достаточно гибкий график позволяли ей возвращаться домой, готовить ужин и быть рядом с детьми.

Во время учёбы на юридическом факультете Университета Лойолы в Чикаго она познакомилась с Джеролдом Альтманом — красивым однокурсником, который был старше её на три года. Джерри происходил из семьи предпринимателя, занимавшегося производством обуви. До знакомства с Конни он учился в Уортонской школе бизнеса Пенсильванского университета, работал консультантом в Бостоне и уже успел однажды жениться. После свадьбы Конни сохранила свою фамилию. Спустя несколько лет семья перебралась из Чикаго в Сент-Луис, родной город обоих супругов.

Джерри занялся недвижимостью и управлением объектами. Какое-то время он был главным юрисконсультом и вице-президентом девелоперской компании Roberts Companies. Он любил людей, интересовался доступным жильём и работал над коммерческими и жилыми проектами, направленными на оживление районов Сент-Луиса и создание местных сообществ. Позднее Сэм часто повторял один из главных уроков отца: «Людям нужно помогать всегда. Даже если кажется, что у тебя совсем нет времени, ты всё равно находишь способ».

У Конни и Джерри один за другим родились трое сыновей: Сэм, Макс, Джек. А через пять лет после Джека — и через девять лет после Сэма — родилась Энни. Конни была счастлива: наконец дочь. Себя Конни называла атеисткой, хотя культурно ощущала связь с еврейством. Джерри был религиознее. Он посещал службы в дни важнейших еврейских праздников и настоял на том, чтобы все четверо детей прошли бар- или бат-мицву. Рациональность и дисциплина матери, а также духовность отца и его убеждение в необходимости служить другим людям по-разному проявятся в каждом из детей.

Сэм с раннего детства отличался редкой целеустремлённостью и почти ненасытным любопытством. В два года он научился пользоваться семейным видеомагнитофоном. В три — уже пытался его чинить. Когда пять лет спустя родители подарили ему компьютер Mac, он быстро освоил программирование и начал разбирать машину на части. Роль старшего брата тоже пришлась ему по душе. Иногда он командовал младшими. Иногда заботился о них. И почти всегда хотел победить. Настольные игры превращались в соревнование, которое Сэм категорически не желал проигрывать.

Он тяжело переносил поражения, но особенно любил вкус победы. Однажды бабушка подарила каждому внуку немного акций. Сэм выбрал Apple. Джек — Applebee's. Через годы это стало семейной шуткой. Стоимость акций Джека почти не изменилась. Акции Сэма взлетели. Джек вспоминал: — Твои Apple выросли настолько, что я даже думать об этом не хочу. В сотни и сотни раз. Сэм с явным удовольствием отвечал: — Да. Прилично выросли.

Когда Сэм стал старше, мать предложила ему перейти в частную школу John Burroughs, известную сильной академической программой и впечатляющим списком знаменитых выпускников. Сэм согласился. Макс тоже попробовал, но позже ушёл. Джек отказался. Энни впоследствии пошла по стопам старшего брата. В Burroughs Сэм чувствовал себя прекрасно. Он хорошо учился, был общительным, уверенным в себе, любил дурачиться и смешить окружающих. Его интересовали не только математика, естественные науки и технологии. Он хорошо писал и участвовал во множестве внеклассных занятий. Возглавлял выпуск школьного ежегодника. Был капитаном команды по водному поло. Участвовал в «Модели ООН», где школьники имитировали работу Организации Объединённых Наций и спорили о международной политике. Его преподаватель английского Энди Эбботт, впоследствии ставший директором школы, вспоминал: — Я тогда думал — и теперь мне даже немного неловко это признавать: «Надеюсь, он не уйдёт в технологии. Он настолько творческий и так хорошо пишет». Мне хотелось, чтобы он стал писателем или кем-то вроде того.

Но уже тогда у Альтмана проявлялась другая способность: он умел вести за собой. И ему нравилось испытывать границы того, что считалось допустимым. В довольно консервативной школе он однажды попал в неприятности после того, как на ежегодном школьном мероприятии организовал со своей командой по водному поло стриптиз — до плавок Speedo. В эти же годы он открыто рассказал родителям и одноклассникам, что он гей. Мать была удивлена, но приняла это. Остальная семья тоже. Однако группа христианских учеников школы отреагировала иначе. В Национальный день каминг-аута они бойкотировали школьное собрание, которое Альтман проводил на тему сексуальности. Сэму было семнадцать. Он решил ответить им публично. Перед всей школой он произнёс речь, которая, по словам его консультанта по поступлению в колледж, серьёзно изменила школьную атмосферу. Последнюю фразу Альтман позже вспоминал так: «Либо у вас есть терпимость, необходимая открытому сообществу, либо её нет. Нельзя выбирать, к кому вы будете терпимы, а к кому — нет».

Но за уверенным внешним обликом скрывался гораздо более чувствительный человек. Сэма сильно занимало, что о нём думают окружающие. Он нередко испытывал тревогу. И эта черта останется с ним на всю жизнь. Позже, когда его известность в Кремниевой долине будет расти, он иногда будет звонить матери из-за обычной головной боли, уже убедив себя, что у него менингит или лимфома. Во время одних переговоров он настолько запаникует, что вынужден будет лечь прямо на пол — без рубашки, раскинув руки, — чтобы успокоиться. Именно сочетание двух черт —

необычайной амбициозности и глубокой чувствительности — во многом определит его карьеру. Журналист *The New Yorker* Тэд Френд, проведший с Альтманом много часов во время подготовки профиля 2016 года, заметит эту двойственность. Почти в любом вопросе Альтман одновременно стремился двигаться вперёд как можно быстрее — и прекрасно чувствовал опасность. Логика могла выглядеть так: создать AGI как можно скорее. И одновременно: только не уничтожить человечество.

После окончания школы в 2003 году Альтман уехал со Среднего Запада в Стэнфордский университет — прежде всего потому, что тот находился в самом сердце технологической индустрии.

Правда, сразу становиться технарём он не собирался. Как и надеялся его школьный преподаватель, Альтман всерьёз думал о писательстве. Очень недолго рассматривал даже карьеру инвестиционного банкира. Но в итоге победила давняя страсть к компьютерам и программированию. Позднее он объяснял: «Я понял, что миру не нужен семимиллионный роман. Не там я мог принести наибольшую пользу. И, кроме того, в подобных областях обычно гораздо труднее заработать много денег — а иногда и просто достаточно денег».

В Стэнфорде Альтман изучал информатику и особенно интересовался искусственным интеллектом и компьютерной безопасностью. Он глубоко погружался в задания. Один однокурсник вспоминал, как Альтман буквально разобрал до низкоуровневого кода программу, которую должен был использовать для домашней работы, — и обнаружил ошибку в самом задании. На втором курсе его заинтересовали мобильные технологии. Узнав, что вскоре почти каждый телефон будет оснащён GPS, он пришёл на студенческое мероприятие для предпринимателей, вышел на сцену с раскладным телефоном и предложил желающим присоединиться к созданию сервиса, использующего геолокацию.

Примерно тогда он познакомился с Полом Грэмом — предпринимателем и влиятельным технологическим блогером, который вместе со своей подругой Джессикой Ливингстон создавал новый стартап-инкубатор: Y Combinator. В 2005 году Альтман попал в самый первый набор YC со своим новым стартапом Loopt. Лето он провёл в Кембридже, штат Массачусетс, где первоначально работал акселератор. Loopt был социальной сетью, использовавшей геолокацию. Приложение сообщало пользователю, если рядом находился друг, или рекомендовало близлежащие рестораны. Альтман работал тем летом настолько много и питался настолько плохо — в основном лапшой быстрого приготовления, — что заработал цингу.

Но не жалел. Позднее он будет советовать молодым предпринимателям: «В начале карьеры работайте очень много. Потом это окупается, как сложный процент». В Стэнфорд Альтман уже не вернулся. К концу 2005 года он и его партнёры вели переговоры с венчурными фондами New Enterprise Associates и Sequoia о финансировании в размере пяти миллионов долларов. Альтман решил рискнуть. И бросил университет. **Loopt так и не стал большим успехом.**

После семи лет работы, в 2012 году, Альтман продал компанию за 43,4 миллиона долларов — примерно за ту же сумму, которую инвесторы успели в неё вложить. Но если бы вы в те годы слушали интервью Альтмана и его инвесторов, могло показаться, что Loopt стоит на пороге великой технологической революции. И именно здесь особенно хорошо видно, откуда впоследствии вырос успех Альтмана.

В ранних интервью он продавал публике вещь куда менее впечатляющую, чем искусственный интеллект: по сути, ещё один сервис геолокации, конкурент Foursquare, который в итоге так и не взлетел. Но Альтман умел превратить даже это в историю о будущем.

Его способность работать с прессой и его талант заключать сделки — два важнейших инструмента будущего восхождения — опирались на одно редкое качество: он превосходно умел рассказывать истории. И это было у него совершенно естественно. Даже если сегодня смотреть старые интервью, уже зная, что Loopt потерпит неудачу, трудно не увлечься тем, что он говорит. Альтман спокойно и уверенно объясняет, почему его компания занимает уникальное место на рынке. Почему она является частью большого, неудержимого движения технологий вперёд. Почему пользователям нужен именно этот сервис. Почему рекламодатели уже готовы за него бороться. И главное: не ставьте против него — успех всё равно неизбежен.

В июне 2010 года, рассказывая технологическому блогеру Роберту Скоблу о новом приложении Loopt Star, Альтман говорил: — Реакция была потрясающей. Приложение позволяло рекламодателям отправлять людям специальные предложения в зависимости от их местонахождения — например скидки в ближайших ресторанах или магазинах. Альтман утверждал, что общество уже прошло важнейший психологический рубеж: — Теперь ценность от того, что я делюсь своим местоположением, уже перевешивает опасения за частную жизнь. Через несколько лет делиться геолокацией станет нормой. А странным будет как раз тот, кто этого не делает.

Несколько месяцев спустя в интервью CNN Business он вообще назвал различие между жизнью в интернете и жизнью «в реальности» бессмысленным. Мобильные устройства, считал Альтман, уже

стирали эту границу. — Весь мир становится мобильным. И весь мир движется к тому, что ваши данные и ваши сервисы будут доступны вам повсюду, где бы вы ни находились.

Даже разговор об опасностях продукта Альтман умел превращать в часть рекламной стратегии. В 2008 году Джессика Лессин, тогда журналист *The Wall Street Journal*, сказала ему, что собирается написать статью о рисках для частной жизни, связанных с отслеживанием местоположения. Альтман не стал защищаться. Он предложил помочь. И отправил ей длинный список рисков, которые команда Loopt уже сама выявила, вместе с предложениями по их устранению. Подтекст был почти очевиден: будущее всё равно будет устроено именно так. Поэтому лучше заранее научиться с ним жить. Позже Лессин сформулировала своё впечатление от Альтмана так: «Он хотел не просто создать стартап. Он хотел написать правила».

КАК СТРОИТСЯ ВЛИЯНИЕ

Именно в годы Loopt Альтман создал связи и отточил навыки, которые впоследствии станут его главным капиталом. В середине 2000-х и начале 2010-х он оказался прямо в эпицентре взрывного роста Кремниевой долины. Новые стартапы появлялись один за другим. Вокруг были амбициозные люди, которые не желали ждать разрешения изменить мир. И Альтман постоянно находился среди них. Когда Loopt только начинал работу, его офис располагался буквально по соседству с молодым стартапом YouTube. В первом наборе Y Combinator вместе с Альтманом были Стив Хаффман и Крис Слоу — соответственно сооснователь и один из первых инженеров Reddit. В 2014 году Альтман войдёт в совет директоров Reddit и со временем будет владеть даже большей долей компании, чем Хаффман. Ещё одним участником того же набора YC был Эмметт Шир, сооснователь Twitch. Почти двадцать лет спустя именно он ненадолго станет временным генеральным директором OpenAI после увольнения Альтмана. Круг замкнётся самым неожиданным образом. Альтман одновременно учился двум вещам: как правильно подавать историю прессе и как заключать сделки, которые кажутся невозможными.

Даже будучи генеральным директором малоизвестного стартапа, он сумел договориться о партнёрствах с крупнейшими американскими операторами мобильной связи. Люди, работавшие с ним, говорили, что его формула складывалась из нескольких качеств. Он великолепно слушал. Был готов помогать. И умел представить собственное предложение именно в тех категориях, которые были важны собеседнику. То есть не просто объяснить: «Вот что у меня есть». А понять: «Чего хотите вы?» — и показать, как его предложение может это дать. Позднее, когда Альтман превратился в одного из богатейших и наиболее влиятельных людей технологической индустрии, этот метод стал ещё эффективнее: теперь ему действительно было что предложить. О нём говорили: «Майкл Джордан среди слушателей». А Джефф Ралстон, который позже возглавил Y Combinator, называл Альтмана:

«Усэйном Болтом в привлечении инвестиций». Ралстон объяснял: — Когда вы привлекаете деньги, по сути вы рассказываете человеку историю о будущем вашего проекта. Историю, в которой этот проект или эта компания становятся невероятно успешными. Сэм умеет рассказать такую историю, в которой вам хочется участвовать. Она увлекательна. Она кажется реальной. Более того — кажется вероятной.

Ралстон сравнивал этот талант со знаменитым «полем искажения реальности» Стива Джобса. Джобс мог настолько убедительно описать будущее, что его версия мира вытесняла всё остальное. И не всегда это было просто иллюзией. Потому что Джобс действительно создавал вещи, которые потом меняли реальность. Ралстон добавлял:

— И Сэм, очевидно, тоже.

Но у этой способности была и другая сторона. Один человек, много лет работавший с Альтманом, описывал её так: — Сэм запоминает мельчайшие подробности о вас. Он невероятно внимателен. Но отчасти он использует это для того, чтобы понять, как именно на вас воздействовать. Он прекрасно

подстраивается под то, что вы говорите. Вам кажется, что разговор движется вперёд, что вы вместе чего-то достигаете. А потом спустя время вы понимаете, что всё это время просто бежали на месте.

И уже в Loopt появились обвинения, которые много лет спустя снова прозвучат во время кризиса в OpenAI. По данным *The Wall Street Journal*, дважды за время руководства Loopt высокопоставленные сотрудники приходили к совету директоров и просили уволить Альтмана. Претензий было две. Первая: он порой действовал прежде всего в собственных интересах, а не в интересах компании — иногда даже во вред компании. Вторая: у него была склонность искажать правду. С этой претензией было сложнее. Иногда речь шла о настолько незначительных деталях, что люди даже не понимали, зачем было лгать. Каждый отдельный случай казался мелочью. Один человек сравнивал их с маленькими порезами от бумаги. Один порез — пустяк. Но когда их становится сотни, результат совсем другой. Со временем в компании возникла атмосфера постоянного недоверия и хаоса.

И всё же кризис закончился так, как впоследствии будет заканчиваться для Альтмана множество других конфликтов: он победил. Совет директоров Loopt встал на его сторону.

Несмотря на скромные результаты самой компании, лично Альтман вышел из истории Loopt гораздо сильнее, чем вошёл в неё. Стартап стал для него трамплином. Через него Альтман проникал во всё более влиятельные круги Кремниевой долины. Затем использовал эти связи, чтобы организовать продажу компании. А продажа сделала его богатым. В двадцать шесть лет Альтман получил от сделки около пяти миллионов долларов. Сам он считал результат разочарованием. Для сравнения: Стив Джобс к двадцати пяти годам уже обладал состоянием примерно в 256 миллионов. Но вскоре денег у Альтмана станет значительно больше. И постепенно они начнут менять его образ жизни. Он перестанет сам ходить за продуктами. Начнёт летать частными самолётами. Станет собирать дорогие спортивные автомобили — среди них McLaren и чрезвычайно редкий Koenigsegg стоимостью около пяти миллионов долларов. Он увлечётся гонками. Какое-то время будет ездить на Burning Man — знаменитый недельный фестиваль в пустыне, известный искусством, психоделической культурой и сексуальной свободой. Как и многие представители кремниеводолинской элиты, Альтман время от времени употреблял кетамин — препарат, который используется как рекреационный наркотик, но может также легально назначаться для лечения депрессии.

Добившись успеха, Альтман потянул за собой и братьев. В 2012 году вместе с Джеком он создал собственный инвестиционный фонд Hydrazine Capital. Джек изучал экономику в Принстоне и сначала пробовал себя в инвестиционном банкинге. Позже он перешёл в технологическую отрасль и основал стартап Lattice. Тот получит финансирование от Y Combinator уже после того, как президентом акселератора станет Сэм. Макс изучал информатику в Duke University, некоторое время работал в Microsoft, затем стал трейдером. В 2014 году он перешёл в другой стартап YC — Zenefits. А ещё через два года присоединился к Сэму и Джеку в Hydrazine Capital. В какой-то момент оба младших брата временно переехали к Сэму. По крайней мере, предполагалось, что временно. В итоге трое братьев проживут вместе долгие годы.

Семья и бизнес окажутся переплетены настолько тесно, что отделить одно от другого станет почти невозможно.

ДВА НАСТАВНИКА

Особенно важными для будущего Альтмана стали отношения с двумя людьми: Полом Грэмом и Питером Тилем. Пол Грэм, которого в Кремниевой долине обычно называли просто PG, сначала прославился как сооснователь Viaweb — стартапа, который Yahoo купила в 1998 году за 49 миллионов долларов. Позднее он стал известным блогером, писавшим о стартапах, предпринимательстве и венчурном капитале. А затем создал Y Combinator. Через YC взгляды Грэма начали влиять на целые поколения предпринимателей. Каждый успешный стартап усиливал престиж акселератора. А вместе с ним — и самого Грэма. К моменту продажи Loopt через YC уже прошли компании, которые стали

или вскоре станут миллиардными: Dropbox, Airbnb и другие. Y Combinator постепенно превратился в самый элитарный клуб Кремниевой долины.

Попал внутрь — и перед тобой открывались двери. У тебя сразу появлялся статус. Доступ к инвесторам. Связи с другими выпускниками YC. Первые потенциальные клиенты. И зачастую более высокая оценка компании. Не попал — ничего этого нет. Грэм стал одним из людей, определявших не только то, какие стартапы получат деньги, но и само представление Кремниевой долины о том, каким должен быть «настоящий основатель». Многие предприниматели воспринимали его взгляды почти как набор законов. Но критиков у него тоже хватало. Грэм любил говорить о технологической индустрии как о меритократии — системе, где побеждает лучший. Однако сам YC он построил как довольно закрытый клуб. Когда его критиковали за небольшое число женщин среди основателей стартапов, он объяснял это тем, что большинство женщин просто с ранних лет не готовили к успеху в технологическом предпринимательстве.

А анализируя собеседования кандидатов YC, Грэм утверждал, что выявил десятки признаков, позволяющих предсказывать неудачу. Среди них был, например, «сильный иностранный акцент». Если у кандидата одновременно обнаруживалось несколько подобных признаков, считал Грэм, вероятность провала возрастала. По его мнению, это было не проявлением предвзятости системы отбора, а объективным сигналом, подтверждённым данными.

В Альтмана Грэм поверил практически сразу. В 2006 году он написал в блоге о своей первой встрече с ним, когда Сэм ещё был студентом второго курса: «Loopt, вероятно, самый многообещающий стартап из всех, что мы пока финансировали. Но Сэм Альтман — совершенно необычный парень. Через три минуты после знакомства я помню, как подумал: “Ага. Наверное, вот таким был Билл Гейтс в девятнадцать лет”». Очень скоро Грэм захотел найти ещё таких Альтманов. Он спросил Сэма: какой вопрос нужно добавить в анкету Y Combinator, чтобы обнаруживать людей, похожих на него?

Альтман предложил вопрос, который вскоре станет одним из самых знаменитых в YC: «Расскажите о случае, когда вам особенно успешно удалось “взломать” какую-либо систему — не компьютерную — в свою пользу». Этот вопрос прекрасно отражал новую предпринимательскую этику, которую YC будет годами воспитывать в своих выпускниках: не обязательно подчиняться правилам. Правила можно обойти. Можно найти лазейку. Можно переписать их. А если потребуется — сломать. Главное — победить.

К двадцати трём годам Альтмана Грэм уже сравнивал со Стивом Джобсом. Он писал: «Когда я консультирую стартапы, чаще всего ссылаюсь на двух основателей — Стива Джобса и Сэма Альтмана. В вопросах дизайна я спрашиваю: “Что сделал бы Стив?” А в вопросах стратегии и амбиций: “Что сделал бы Сэма?”» «Сэма» — так звучало прозвище Альтмана и его имя в X. Именно почти безграничная вера Пола Грэма в Альтмана в конечном счёте приведёт того к вершине Y Combinator. В 2014 году, через два года после продажи Loopt, двадцативосьмилетний Альтман станет президентом YC. Когда Грэм на кухне спросил, не хочет ли Сэм стать его преемником, тот не смог сдержать улыбку. Альтман позднее объяснял привлекательность должности очень просто: «Y Combinator в определённой степени получает возможность направлять развитие технологий». А Пол Грэм говорил о нём ещё грандиознее: «Мне кажется, его цель — создать всё будущее». Историю о назначении будут повторять столько раз, что она превратится в одну из легенд Кремниевой долины. Один бывший основатель YC шутил: — Если Сэм улыбается, то обычно делает это совершенно сознательно. Но один раз он действительно не мог остановиться — когда PG предложил ему возглавить YC.

Для многих выбор Грэма оказался неожиданностью. Но у самого Грэма сомнений не было. Джессика Ливингстон позднее вспоминала: «Не существовало списка кандидатов, где Сэм стоял бы первым. Был просто Сэм». Вторым важнейшим наставником Альтмана стал **Питер Тиль**. Это был ещё один человек огромного влияния в технологическом мире. Тиль стал миллиардером как сооснователь PayPal и компании по анализу данных Palantir, а также как один из первых инвесторов Facebook. Как и Пол Грэм, он был фигурой далеко не бесспорной. В 2016 году Тиль оказался одним из немногих крупных представителей технологической индустрии, кто открыто поддержал Дональда Трампа.

Кроме того, он тайно финансировал судебный процесс, который в итоге уничтожил Gawker Media, — в отместку за то, что сайт почти десятью годами ранее публично раскрыл его гомосексуальность.

После продажи Loopt Альтман переживал тяжёлый период. Он расстался и с одним из сооснователей компании, и с человеком, который девять лет был его партнёром. Разбитый личной драмой и не вполне понимающий, куда двигаться дальше в профессиональном плане, Альтман взял годичный перерыв, создал инвестиционный фонд Hydrazine Capital и привлёк в него **21 миллион долларов**. Большую часть этих денег вложил Питер Тиль, почти на двадцать лет старше Альтмана. Когда Альтман стал партнёром Y Combinator, он начал использовать Hydrazine для инвестиций в компании акселератора, одновременно помогая венчурному фонду Тиля Founders Fund находить наиболее перспективные проекты.

Состояние Тиля за это время выросло многократно. Их отношения стали чрезвычайно близкими. Однажды их даже сравнили только с одной другой парой наставник — ученик: отношениями самого Тиля с Марком Цукербергом.

Пол Грэм и Питер Тиль во многом сформировали мировоззрение Альтмана. Они повлияли на то, как он строил компании. Как понимал власть. Как вёл политическую игру. Оба внушали ему одну важнейшую мысль: **масштаб имеет решающее значение.** И ещё одну: **капитализм работает эффективнее государства.** В 2013 году Альтман написал в своём блоге: «Первый совет о стартапах, который я когда-либо услышал, звучал так: “Увеличение продаж решит все проблемы”». Пост назывался **«Рост и государство»**, и в нём Альтман прямо благодарил Грэма и Тиля за влияние на свои взгляды. Он пояснял, что по сути это всего лишь другая формулировка мысли Пола Грэма: **рост — критически важен.** Для стартапа больше продаж означает больше денег. Больше денег — возможность нанять лучших людей.

Лучшие люди — меньше внутренних конфликтов и больше шансов победить. Но Альтман переносил ту же логику и на целые страны. Больше экономического роста — значит больше технологических инноваций, а значит выше качество жизни. Он считал, что дисфункция американского правительства угрожает этому циклу. Его формула была жёсткой: «Либо ты растёшь, либо медленно умираешь». И, по его мнению, американское государство как раз умирало. Альтман писал, что без экономического роста демократия перестаёт нормально работать, потому что избиратели оказываются в системе с нулевой суммой: если кому-то становится лучше, другой начинает думать, что у него что-то отняли.

Со временем эта идея превратилась в одну из главных тем всей карьеры Альтмана. В 2017 году он говорил: — Лучшее, что частный сектор может сделать для того, чтобы страна снова встала на правильный путь, — вернуть экономический рост.

Он рассуждал, что Соединённые Штаты пережили около двухсот лет почти беспрецедентного роста. Сто лет территориального расширения. Затем сто лет бурного технологического развития. При этом Альтман почти не касался кровавой истории колониальной экспансии и сложных последствий неограниченной индустриализации для работников и окружающей среды.

Он просто подводил итог: — И люди в основном были довольно счастливы. А теперь — нет. В 2019 году он формулировал мысль ещё яснее: «Устойчивый экономический рост почти всегда является моральным благом». И добавлял: — Отчасти именно это мотивирует меня работать над Y Combinator и OpenAI. Я хочу вернуть нас к устойчивому экономическому росту, к миру, где жизнь большинства людей каждый год становится лучше и где мы снова ощущаем общий дух успеха.

В строительстве компаний Альтман часто опирался и на другую идею Тиля: **нужно стремиться к монополии.** Тиль считал, что настоящий успешный основатель должен не просто хорошо конкурировать. Он должен сделать так, чтобы конкурировать с ним было почти невозможно. В 2014 году Альтман вернулся в Стэнфорд — теперь уже как преподаватель курса **How to Start a Startup**. Одну из лекций он отдал Питеру Тилю. Она называлась вызывающе: **«Конкуренция — для неудачников».** Тиль объяснял: монополии хороши потому, что такие компании устойчивее, живут

дольше, располагают большим капиталом и обычно являются признаком того, что было создано нечто действительно ценное.

Для построения монополии, по его мнению, нужны несколько вещей: собственная технология, сетевые эффекты, экономия от масштаба, сильный бренд. Но главное — всё это должно сохранять ценность годами. Особенно важно было не потерять технологическое лидерство. Тиль предупреждал: — Вы не хотите, чтобы кто-то другой вас обошёл.

Он приводил пример производства дисковых накопителей в 1980-е годы. Технология стремительно улучшалась. Каждые пару лет появлялся новый прорыв.

Но каждый раз его делала уже другая компания. Для потребителей это было прекрасно. А вот для владельцев компаний — совсем нет. По мнению Тиля, мало совершить **первый большой прорыв**. Нужно добиться ещё и **последнего прорыва** — такого, который позволит сохранить лидерство надолго. И дальше совершенствовать продукт достаточно быстро, чтобы никто никогда не догнал. Вывод Тиля был почти циничным: — Если будущее устроено так, что инноваций много и другие люди постоянно придумывают что-то новое в вашей области, для общества это прекрасно. Но для вашего бизнеса — не очень.

СЕТЬ КАК ФОРМА ВЛАСТИ

От Грэма и Тиля Альтман усвоил ещё один принцип: **связи — это тоже капитал.** В 2013 году он писал: «Я слышал много разных теорий о том, как в мире делаются дела. Вот лучшая: сочетание сосредоточенности и личных связей». Эту мысль, по словам Альтмана, Чарли Роуз когда-то высказал Полу Грэму, а Грэм передал её ему. Позже Альтман добавил третью составляющую: **веру в себя.** Он говорил: — Для стартапов это действительно важно. Вы должны всерьёз верить, что у вас может получиться. Этой формуле Альтман начал следовать почти религиозно. Он строил отношения с огромной дисциплиной. Сначала предлагал людям своё время и советы. Затем — по мере того как в его распоряжении оказывалось всё больше капитала — начал давать деньги. В этом смысле Тиль был для него идеальным образцом. Тиль годами использовал советы и инвестиции для построения сети. А затем использовал эту сеть, чтобы получать ещё больше связей и ещё больше денег. Молодым предпринимателям, которых хотел втянуть в свою орбиту, он давал небольшие суммы, наставничество и доступ к нужным людям. Альтман научился делать то же самое. И чем богаче и влиятельнее он становился, тем масштабнее использовал этот подход.

Он регулярно устраивал у себя дома ужины и встречи для самых разных, но пересекающихся кругов людей. Там были: основатели из Y Combinator, предприниматели из компаний, в которые он инвестировал, люди с похожими инвестиционными интересами, представители растущего сообщества предпринимателей-геев. Советы он часто давал короткими звонками и сообщениями. Иногда разговор длился всего две минуты. Ему было важно вместить как можно больше контактов в один день. Он мог познакомить двух людей по электронной почте одним словом: **«meet» — «познакомьтесь».** Или вообще одним знаком вопроса: **«?»** Такой стиль, к слову, был знаменитой привычкой Джеффа Безоса. В инвестициях Альтман тоже предпочитал не несколько огромных ставок, а множество небольших. По оценке *The Wall Street Journal* за июнь 2024 года, через YC, Hydrazine и другие структуры он приобрёл финансовые связи более чем с **четырьмя сотнями компаний**.

Среди близкого окружения Альтмана становилось трудно найти человека, с которым у него не было бы какого-нибудь финансового пересечения. Второй стартап, в который он когда-либо инвестировал, оказался и одним из самых удачных: **Stripe**. Её первым техническим директором был Грег Брокман. Альтман рано вложился в Airbnb. Её сооснователь и генеральный директор Брайан Чески стал одним из его ближайших друзей. Он инвестировал в биотехнологическую компанию Conception, созданную его бывшим партнёром и другом Мэттом Крисилоффом. В других сделках выступал совместным инвестором с ещё одним бывшим партнёром, венчурным капиталистом Лаки

Грумом. Люди вокруг Альтмана воспринимали это как проявление исключительной щедрости. Он действительно часто делился ресурсами. Предоставлял свои дома друзьям и знакомым. Давал деньги. Помогал даже совершенно незнакомым людям. Один человек вспоминал, как Альтман отправил деньги жителю Эфиопии, который просто написал ему по электронной почте и попросил помочь купить ноутбук. Во время кризиса Silicon Valley Bank в 2023 году, когда крупнейшее банковское банкротство со времён 2008 года поставило под угрозу множество стартапов, Альтман переводил деньги компаниям практически без документов, чтобы те не закрывались и не увольняли людей. Лаки Грум говорил: — Это чрезвычайно редкое качество. Щедрость Сэма действительно повлияла и на меня. Я очень ему за это благодарен.

ПОЛИТИКА

Тот же принцип Альтман начал применять и к политикам. Отчасти снова по примеру Тиля. Но политически они двигались в разные стороны. Тиль использовал свои деньги для поддержки республиканцев, вкладывая в их кампании десятки миллионов долларов. Альтман, наоборот, всё активнее помогал демократам. Организовывал сборы средств. Выписывал чеки. На некоторое время политические разногласия даже ухудшили отношения между ним и Тилем. В 2017 году Альтман решил лучше понять противоположный лагерь. Он отправился в путешествие по Америке — чем-то напоминавшее поездку другого протеже Тиля, Марка Цукерберга, — и лично поговорил со ста сторонниками Трампа. Сам Альтман в то время даже размышлял о политической карьере. Например, о выдвижении на пост губернатора Калифорнии. Логика была прагматичной: губернатор Калифорнии фактически руководит экономикой, которая была бы пятой по величине в мире. Это могло стать хорошей точкой входа для исправления того, что Альтман считал дисфункцией американской политической системы.

Он даже опубликовал политический манифест под названием **The United Slate**. В нём было три основных принципа: процветание благодаря технологиям; экономическая справедливость; личная свобода. Альтман организовывал фокус-группы, чтобы проверить, как избиратели воспринимают его потенциальную кандидатуру. Близкие шутили: может, сразу идти в президенты США? В итоге политиком он не стал. Фокус-группы решили, что он выглядит слишком молодо. Но постепенно он всё больше **начинал вести себя как политик**.

В первые годы руководства Y Combinator Альтман всё ещё выглядел почти мальчишкой.

Круглые щёки.

Один-единственный пиджак.

Он мог сидеть, закинув ногу на стул, или буквально устроиться на кресле как птица на жердочке.

Говорил резко.

Любил гиперболы.

Часто ругался.

Легко упоминал провокационные детали своей личной жизни — например коллекцию оружия.

Быстро раздражался.

Особенно если считал людей неэффективными.

Однажды он даже написал программу, которая оценивала основателей стартапов YC по тому, **как быстро те отвечают на электронные письма**.

Но за несколько лет образ изменился.

Альтман отшлифовал внешность.

Футболки и шорты-карго сменились облегающими рубашками Henley и джинсами.

За один год он набрал около восьми килограммов мышц, чтобы его небольшая фигура выглядела солиднее.

Он научился меньше говорить.

Больше спрашивать.

Хмурить брови и создавать впечатление вдумчивой скромности.

С близкими друзьями и в частной обстановке раздражение и вспышки гнева никуда не исчезли.

Но публично он становился воплощением **приятного, мягкого человека**.

Щедро хвалил других.

Писал сообщения исключительно строчными буквами.

Использовал много улыбающихся и грустных смайликов.

Раздавал сотрудникам личный номер телефона и говорил, что они могут писать в любое время.

И действительно внимательно отвечал на отзывы.

Он старался не показывать отрицательных эмоций.

Избегал прямых конфликтов.

Избегал говорить людям «нет».

Один бывший сотрудник OpenAI вспоминал, что после его увольнения Альтман лично связался с ним и предложил в качестве утешения кетамин и алкоголь.

Тот шутил:

— Мне кажется, любые отношения с Сэмом заканчиваются хорошо — хотите вы этого или нет.

ЧЕЛОВЕК-ИНСТИТУТ

Постепенно Альтман сам превратился в институт. Y Combinator был для него платформой и ускорителем. Власть YC становилась его личной властью. Связи YC — его связями. А личная сеть и репутация постепенно превратились в его главную валюту. Он регулярно встречался с политиками, которые видели в нём посредника между Вашингтоном и Кремниевой долиной. В 2016 году к нему за советом обращался министр обороны при Бараке Обаме Эштон Картер. Его интересовало, как Министерству обороны привлекать молодых талантливых специалистов из технологической индустрии. Три года спустя, в марте 2019 года, в день ухода Альтмана из Y Combinator, к нему приехал уже Чак Шумер. Тогда он был лидером демократического меньшинства в Сенате США. Визит в OpenAI проходил почти тайно, в сопровождении охраны. Шумер сидел рядом с Альтманом в кресле на фоне телевизора, где горел виртуальный камин, и говорил сотрудникам: — Вы занимаетесь важной работой. Затем добавил: — Мы пока не до конца понимаем, что именно вы делаете, но это важно. И заключил: — Я знаю Сэма. Вы в надёжных руках.

Но чем выше поднимался Альтман, тем больше появлялось людей, которые воспринимали его совсем иначе. Повторялись те же обвинения, которые когда-то звучали в Loopt: **он стремится к власти прежде всего ради себя** и **слишком легко обращается с правдой.** Многие, кто получил от Альтмана деньги, советы и доступ к его сети, становились преданными сторонниками. Но другие видели в нём человека, который почти дьявольски хорошо умеет повернуть любую ситуацию в свою пользу. Для равных ему — партнёров по YC и других влиятельных людей — это могло быть просто раздражающей особенностью. Для рядовых сотрудников и людей, обладавших гораздо меньшей властью, — уже источником страха. А для тех, кто категорически не разделял его взгляды на будущее, Альтман превращался в настоящую угрозу.

ЭННИ

Восхождение Альтмана имело и ещё одну болезненную сторону: оно разрушило его отношения с младшей сестрой Энни. В детстве и даже когда Сэму было уже за двадцать, они оставались близки. Она была самой младшей. Он — самым старшим. Он защищал её. Энни интересовалась наукой, но одновременно была более творческой и эмоциональной из детей. Сэм иногда просил её высказать мнение о людях, с которыми встречался. Делился внутренними страхами и переживаниями. Но чем глубже он погружался в мир Кремниевой долины, тем сильнее Энни чувствовала, что брат возводит

вокруг своей чувствительной стороны всё более толстую стену. Она вспоминала, что он даже рассказывал ей о новых психологических приёмах, которым научился. Например: писать в электронных письмах **меньше слов**, чтобы производить впечатление более влиятельного руководителя. Сначала Энни было просто грустно.

Потом стало страшно. Она начала сомневаться, осталась ли в брате вообще та мягкая и ранимая часть, которую она знала раньше. Она говорила: «Какое-то время я всё ещё иногда её замечала. Поэтому и сохраняла близость. А потом человеком, которому он начал причинять вред, стала уже я».

Когда Сэм только начал богатеть, по словам Энни, его тогдашний партнёр ввёл для него правило: на каждую крупную личную покупку нужно жертвовать такую же сумму на благотворительность. Какое-то время это удерживало рост его роскошного образа жизни. Но по мере того как деньги начали приходить быстрее, чем он мог их тратить, Энни стала воспринимать отношения брата с богатством всё более тревожно. Ей казалось, что он начинает накапливать деньги и всё хуже понимает людей, которые живут в нужде. С конца 2019 года и в первой половине 2020-го, по предоставленной автору переписке, Энни несколько раз просила Сэма и других членов семьи о срочной финансовой помощи — в том числе для оплаты жилья и медицинских расходов. Иногда ей отказывали, иногда помогать не хотели. В тот период она переживала тяжёлые физические и психологические проблемы, что подтверждалось её медицинскими записями. Ситуацию усугубила внезапная смерть их отца. Энни столкнулась с нестабильным жильём. В отчаянии, пытаясь заработать на жизнь, она занялась секс-работой. Летом 2020 года, как раз когда OpenAI под руководством Сэма начала получать первую большую волну общественного внимания, Энни полностью прекратила общение с семьёй.

Автор подчёркивает, что у этой истории есть и другая сторона. Вполне возможно, что Сэм, его братья и их мать пытались подтолкнуть Энни к финансовой самостоятельности. Это сложная и болезненная семейная ситуация, которую трудно объективно оценивать по неполной информации. В январе 2025 года Сэм, его мать и два брата выпустили публичное заявление. Они выразили любовь и тревогу за Энни, но назвали все её обвинения **«совершенно не соответствующими действительности»**. Мать, Конни Гибстайн, предоставила автору более короткую версию того же заявления и отказалась от дальнейших комментариев. Сэм через пресс-службу OpenAI и его братья на подробные вопросы автора не ответили.

Тем не менее история Энни, по мнению автора, удивительным образом перекликается со многими главными темами этой книги. Разрыв между теми, кто выигрывает от прогресса, и теми, кого он оставляет позади. Потеря голоса и возможности влиять на собственную судьбу. Опасность сосредоточения огромной власти не просто в руках компаний, а в руках отдельных людей — без достаточно сильной системы сдержек и противовесов. Со временем действия Энни стали частью и истории самой OpenAI. В 2021 году она решила публично выдвинуть против Сэма тяжёлые обвинения, заявив, что в детстве он подвергал её сексуальному насилию. Семья Альтмана назвала это самым тяжёлым из её «ложных» обвинений. Энни также утверждала, что брат и остальные родственники оставили её без поддержки в самый уязвимый период её жизни. 6 января 2025 года, за два дня до своего тридцать первого дня рождения, она подала против Сэма иск, чтобы уложиться в установленный в Миссури срок давности по таким делам. Её настойчивые попытки публично рассказывать собственную версию событий неизбежно влияли и на Сэма, и на других руководителей OpenAI — особенно в тот момент, когда влияние компании стремительно становилось глобальным.

КАК ВСЁ ЭТО ПРИВЕЛО К OPENAI

Все эти части складываются в одну картину: восхождение Сэма, его характер, его отношения, люди, которых он привлекал — и отталкивал, потоки денег, потоки влияния, борьба за власть. И всё это помогает понять, как стало возможным его внезапное увольнение из OpenAI — пусть оно и

продлилось совсем недолго. На несколько дней весь мир получил редкую возможность заглянуть внутрь борьбы, которая идёт на самом верху за право определять будущее искусственного интеллекта. И увиденное показало: будущее технологии, которая уже перестраивает общество и буквально меняет планету, зависит вовсе не только от науки и инженерии. Оно зависит от **ценностей, амбиций, конфликтующих характеров и человеческих слабостей очень небольшой группы людей**. Людей влиятельных. Людей чрезвычайно богатых. Но всё равно — **людей, способных ошибаться.**

ГЛАВА 2.

ЦИВИЛИЗАТОРСКАЯ МИССИЯ

Первым, кто окончательно решил присоединиться к созданию OpenAI, стал Грег Брокман. В качестве его сооснователя Альтман лично выбрал Илью Суцкевера — исследователя искусственного интеллекта из Google. Именно Альтман неожиданно написал ему по электронной почте и пригласил на тот самый ужин в Rosewood летом 2015 года. Узнав, что там будет Илон Маск, Суцкевер сразу согласился.

Брокман и Суцкевер представляли собой необычную пару. Брокман — высокий, крепкий, доброжелательный — был инженером и человеком стартапов, во многом похожим на Альтмана. Он вырос на небольшой ферме в Северной Дакоте. Между дойкой коров увлёкся математикой, затем наукой. В 2008 году поступил в Гарвард, через два года перевёлся в MIT, а ещё через семестр совсем бросил университет: ему было трудно продолжать учиться, когда за пределами аудитории существовал настоящий мир, где можно было создавать реальные продукты. Он переехал в район Сан-Франциско и присоединился к тогда ещё крошечному стартапу Stripe, где кроме него работали всего три человека. Основателей настолько поразил его талант программиста, что Брокман стал техническим директором. За следующие пять лет он участвовал в создании многих первых продуктов Stripe и помог компании превратиться в одного из гигантов финансовых технологий, обслуживающего цифровые платежи таких компаний, как Amazon и Shopify.

Эти годы принесли Брокману и значительное состояние, и редкий для Кремниевой долины статус человека, который лично участвовал в создании многомиллиардной компании. Суцкевер был другим. Он был прежде всего учёным. Худой, жилистый, родился в Советском Союзе, вырос в Израиле, где очень рано проявил необычайные математические способности. Учителям становилось всё труднее успевать за ним, поэтому родители ещё в восьмом классе записали его на курсы Открытого университета Израиля. В шестнадцать лет он переехал в Торонто. В обычной школе проучился всего месяц. В 2003 году его сразу приняли в Университет Торонто — причём на третий курс бакалавриата. Именно там Суцкевер встретил Джеффри Хинтона, британско-канадского профессора, чьи исследования стали фундаментальными для современной истории искусственного интеллекта. Хинтон был единственным человеком, которого Суцкевер впоследствии называл своим наставником. И влияние Хинтона на его жизнь и работу оказалось огромным. В 2012 году Суцкевер, Хинтон и ещё один аспирант Хинтона — Алекс Крижевский — потрясли весь мир искусственного интеллекта. Они участвовали в конкурсе ImageNet, где команды создавали программы, автоматически распознающие объекты на фотографиях.

Другим участникам никак не удавалось снизить долю ошибок ниже 25 процентов. Команда Хинтона, Суцкевера и Крижевского добилась **15 процентов**. Это был разгром. В начале следующего года Google купила их только что созданную компанию DNNresearch после ожесточённого аукциона за **44 миллиона долларов**. Трое учёных мгновенно стали мультимиллионерами. Но важнее было другое: именно эта сделка стала одним из первых мощных сигналов к массовой коммерциализации искусственного интеллекта. Хинтон вспоминал: — Нам казалось, будто мы оказались в кино.

Брокман и Суцкевер впервые встретились именно на ужине в Rosewood. Другие гости тоже делились на две большие группы: предприниматели и учёные. Разговор постоянно перескакивал от академических споров о направлениях исследований к теме, которая особенно волновала Маска: **не поздно ли ещё обогнать DeepMind и Google?** Для собравшихся это было не просто соревнование компаний. Они считали, что победа над Google необходима, чтобы изменить само направление развития искусственного интеллекта. И довольно быстро все согласились: главный дефицит — **люди**. Лучшие исследователи ИИ уже работали либо в Google, как Суцкевер, либо в других технологических

гигантах. Они получали огромные зарплаты, щедрые льготы и надёжность, которую молодая лаборатория почти не могла предложить.

В центре разговора стоял и AGI — искусственный общий интеллект. В 2015 году само серьёзное обсуждение этой темы было необычным. Большинство уважаемых учёных считали полноценное цифровое воспроизведение человеческого интеллекта научной фантастикой. Или, в лучшем случае, технологией далёкого будущего — через десятилетия. Разговоры о том, что AGI уже настолько близок, что в него пора инвестировать огромные деньги, часто воспринимались как псевдонаука. Демис Хассабис, однако, уже использовал термин AGI для описания целей DeepMind. Причём даже некоторые его собственные сотрудники считали это слишком вызывающим маркетингом. Но участники ужина в Rosewood решили: если они хотят создать организацию, способную бросить DeepMind прямой вызов, им нужен столь же грандиозный ориентир. И AGI лучше всего выражал их амбиции. Даже Суцкевер, который в глубине души считал AGI вполне возможным и позднее станет одним из самых убеждённых сторонников идеи его быстрого появления, сначала чувствовал себя неуютно.

Он боялся: если другие учёные узнают, что он всерьёз обсуждает создание AGI, его научная репутация может пострадать. Брокмана такие опасения не мучили. Он пришёл в ИИ извне. И искренне верил: если приложить достаточно сил и сосредоточиться на цели, AGI может оказаться гораздо ближе, чем все думают. Сам факт, что Суцкевер и другие исследователи хотя бы в частном порядке готовы обсуждать создание новой лаборатории, только укрепил его уверенность. Когда в ту ночь Альтман вёз Брокмана из отеля обратно в Сан-Франциско, тот сказал: **он готов посвятить себя этому проекту.**

В Кремниевой долине любят говорить: **создать компанию вместе — почти то же самое, что вступить в брак.** Поэтому Брокман и Суцкевер сначала словно присматривались друг к другу. Альтман дипломатично объяснил Брокману: ему нужен партнёр, который действительно понимает исследования в области ИИ. Через несколько недель после ужина они встретились вдвоём за ужином в Маунтин-Вью. Оказалось, что совпадение почти идеальное. Позднее Брокман написал: «Хотя мы только познакомились, я уже знал, что всё получится. У нас с Ильёй возникло общение невероятной плотности. Наши идеи усиливали и дополняли друг друга».

Пока Суцкевер обдумывал предложение, Брокман занялся главным: **собирал команду.** Он звонил ведущим людям отрасли и спрашивал, кого они считают лучшими специалистами. Встречался с профессорами и выяснял, кто их самые сильные студенты. Перед разговором с каждым кандидатом тщательно изучал его работы, чтобы понимать, чем именно можно заинтересовать человека. Позднее Альтман посвятил этому отдельную запись в блоге — просто под названием **«Greg»**. Он писал: «Меня часто спрашивают, каким должен быть идеальный сооснователь. Теперь у меня есть ответ: Грег Брокман».

Альтман и Маск тоже активно участвовали в наборе людей. Главной задачей было постепенно убедить исследователей, которым идея AGI казалась слишком смелой. Профессор Беркли Питер Аббил вспоминал, как Маск говорил ему примерно следующее: — Возможно, AGI ещё очень далеко. Но что, если нет? Что, если есть хотя бы один процент вероятности — или даже одна десятая процента, — что это произойдёт в ближайшие пять-десять лет? Разве мы не обязаны отнестись к этому очень серьёзно?

Позднее Аббил стал научным консультантом OpenAI, а затем пришёл туда на постоянную работу вместе с несколькими своими аспирантами. Поначалу многие кандидаты соглашались только при условии, что присоединятся и другие. Брокмана это не остановило. Он выбрал десять инженеров и исследователей, которых хотел получить больше всего, и пригласил всех в Напа-Вэлли. Формально — на вино и разговор.

На самом деле — на групповую вербовку. Он арендовал автобус, чтобы отвезти всех туда и обратно. И даже в дороге, больше часа в каждую сторону, продолжал убеждать их присоединиться. Через три недели, к установленному Брокманом сроку, почти все приняли предложение.

Пока Маск, Альтман и Брокман обсуждали, как представить OpenAI миру, один вопрос занимал их особенно: **какое впечатление должна производить организация?** Они поддержали идею Альтмана сделать OpenAI некоммерческой. И решили максимально подчёркивать то, что было заложено в названии: **открытость.** OpenAI должна была стать **анти-Google**. Она будет вести исследования для всех. Открывать исходный код. Делиться научными результатами. Быть образцом прозрачности. В ноябре 2015 года Брокман написал Маску и Альтману: «Надеюсь, мы войдём в эту область как нейтральная группа, которая стремится широко сотрудничать и изменить сам разговор: не о победе какой-то конкретной компании или группы, а о победе всего человечества».

И добавил в скобках: «Думаю, именно так мы быстрее всего сможем стать ведущим исследовательским институтом». Маск согласился, но думал ещё и о публичном эффекте. Он написал: «Очень важно, чтобы общество болело за наш успех». И предложил сделать заявление об основании OpenAI более масштабным. Особенно его беспокоила сумма. Сто миллионов долларов, считал Маск, выглядели смешно на фоне того, сколько Google и Facebook тратили на ИИ. Поэтому он предложил: **объявить о миллиардном обязательстве по финансированию.** Маск написал: «Это реально. Всё, чего не дадут другие, покрою я».

Но уже в личной переписке основатели признавали: обещанную открытость, возможно, придётся со временем сократить. Например, чтобы опасные технологии не попали в плохие руки. В январе 2016 года, вскоре после запуска OpenAI, Суцкевер написал: «Когда мы приблизимся к созданию ИИ, вероятно, станет разумно быть менее открытыми». И объяснил: «Open в OpenAI означает, что после создания ИИ все должны получить пользу от его результатов. Но это совершенно не означает, что мы обязаны раскрывать всю науку». Маск ответил одним словом: **«Да».** Так уже в первые месяцы существования OpenAI возникло напряжение, которое позднее станет принципиальным: **открытость была одновременно идеалом и инструментом.**

Официально OpenAI представили публике в декабре 2015 года. Заявление специально выпустили вечером в пятницу — одновременно с крупнейшей ежегодной конференцией по искусственному интеллекту Neural Information Processing Systems. Именно там тремя годами раньше Хинтон и Суцкевер фактически устроили аукцион по продаже DNNresearch. Пост назывался: **«Представляем OpenAI».** В нём были перечислены девять основателей. Брокман становился техническим директором. Суцкевер — руководителем исследований.

Маск и Альтман — сопредседателями. Альтман и Брокман присоединились к обещанию Маска обеспечить лаборатории финансирование в размере **миллиарда долларов**. То же сделали Джессика Ливингстон, Питер Тиль и сооснователь LinkedIn Рид Хоффман. Хоффман когда-то работал вместе с Маском и Тилем в PayPal, а затем регулярно инвестировал вместе с другими представителями так называемой **«мафии PayPal»**. При этом в объявлении осторожно уточнялось: «В ближайшие несколько лет мы рассчитываем потратить лишь небольшую часть этой суммы».

Перед самым запуском всё едва не сорвалось из-за Суцкевера. Остальным основателям OpenAI предлагала базовую зарплату в **175 тысяч долларов** и акции YC или SpaceX. Суцкеверу предложили почти **два миллиона долларов** в год. Для некоммерческой организации это была огромная сумма. Но Google предложила ещё больше. Потом подняла предложение снова. В какой-то момент речь шла о зарплате в два-три раза выше. Маск и Альтман несколько раз откладывали официальное объявление, пока Суцкевер мучительно выбирал. Он звонил родителям.

Слушал уговоры Маска. Слушал Брокмана. И в итоге решил уйти из Google. Парадоксально, но именно невероятно щедрое предложение Google убедило его, что OpenAI действительно нужна. Если лучшие исследователи всегда будут принадлежать только богатейшим корпорациям, кто тогда сможет

действовать в интересах человечества? В тот день Маск отправил команде торжественное письмо. Главная задача, написал он, — **нанять лучших людей.** Он обещал помогать в этом всем, чем сможет. И добавил: «Нас численно и ресурсно превосходят с нелепо огромным отрывом организации, которые вы прекрасно знаете. Но правда на нашей стороне. А это многое значит. Мне нравятся наши шансы». Чтобы никто из конкурентов не переманил новых сотрудников, OpenAI немедленно повысила всем базовые зарплаты ещё на **100 тысяч долларов**.

Позднее Маск рассказывал, что Ларри Пейдж пришёл в ярость из-за того, что он лично переманил Суцкевера. После этого отношения между ними почти прекратились — тем более что взгляды на будущее ИИ расходились всё сильнее. Но для OpenAI рекрутинг неожиданно стал проще. В Кремниевой долине грандиозные проекты и имена миллиардеров действовали почти магически. Всего за несколько месяцев число сотрудников удвоилось.

Однако уже тогда в OpenAI проявлялись типичные черты Кремниевой долины на пике её уверенности в себе: размытые цели, огромные деньги, поклонение миллиардерам. И всё же организация очень удачно выбрала позицию. С одной стороны, она воплощала традиционную веру Кремниевой долины в технологию как универсальное решение. С другой — пыталась выглядеть более ответственной альтернативой. И это оказалось своевременно.

После стремительного политического взлёта Дональда Трампа и его победы на выборах 2016 года многие сотрудники технологической индустрии, в основном настроенные либерально, впервые всерьёз задумались о собственной роли в обществе. Началась волна критики Big Tech. Скандалы потрясали Meta и Google. Исследователи ИИ тоже начали задавать вопрос: **не слишком ли быстро мы связали развитие технологии с интересами корпораций?**

К тому времени последствия коммерциализации ИИ становились всё тревожнее. Автоматизированные системы, которые продавались полиции, ипотечным компаниям и кредиторам, усиливали уже существующую дискриминацию по расе, полу и социальному положению.

Алгоритмы Facebook и рекомендательные системы YouTube, по всей видимости, усиливали политическую поляризацию. Распространяли дезинформацию. Подталкивали к экстремизму. Создавали возможности для вмешательства в выборы. А в самом страшном случае Facebook оказался связан с событиями, которые способствовали этническим чисткам в Мьянме. Но и главный альтернативный источник денег — государство — был этически далеко не безупречен. В 2018 году тысячи сотрудников Google протестовали против секретного контракта компании с Пентагоном в рамках Project Maven. Проект предполагал использование ИИ в военных беспилотниках наблюдения. Сотрудники боялись, что это станет основой для автономного оружия. Пентагон тогда отрицал такие планы, хотя позже, во время войны в Украине, его позиция изменится. Среди исследователей ИИ всё чаще звучала циничная формула: **либо продавайся Big Tech, либо военно-промышленному комплексу — либо уходи из исследований.** На этом фоне OpenAI выглядела третьим путём. Не корпорация, подчинённая прибыли. Не государственный военный проект. Один инженер машинного обучения, Чип Хьюен, называл её: **«маяком надежды».**

Но далеко не все были впечатлены. В ночь запуска OpenAI Тимнит Гебру — та самая исследовательница, которая когда-то спросила Маска, почему он так боится ИИ, но не изменения климата, — увидела новость и не могла поверить своим глазам. Всю неделю она находилась на конференции NIPS.

Для неё это был первый раз среди огромного сообщества исследователей ИИ. Гебру родилась в Эфиопии, была эритрейской беженкой и подростком переехала в США. На конференции она снова и снова ощущала, насколько дорого обходится быть чернокожей женщиной в среде, где почти все — белые мужчины. Она была одной из немногих чернокожих участниц. На следующей конференции, в 2016 году, она специально посчитает:

среди **8500 участников** окажется всего шесть других чернокожих исследователей. В Стэнфорде она раньше шутила:

можно создать группу Black in AI — и проводить встречи в одиночку. Она даже представляла юмористический YouTube-канал, где меняла бы причёски и играла разные роли, изображая разговор самой с собой. Но за шутками скрывалась настоящая изоляция.

И в 2015 году конференция напомнила ей, как быстро эта среда может стать враждебной.

На одной вечеринке она подошла за водой. Группа пьяных мужчин в футболках Google Research заметила её и решила развлечься. Они окружили Гебру. Один насильно обнял. Другой поцеловал в щёку и одновременно сделал унизительную фотографию.

На той же конференции профессор домогался до её подруги. И теперь Гебру видела перед собой другую картину:

группа людей, девять из одиннадцати — белые мужчины, получает невиданные ранее суммы денег и рассуждает о теоретической опасности сверхразума, который однажды может захватить мир. А их решение проблемы — **построить собственный, более хороший сверхразум.** Гебру написала резкую критику в форме анонимного открытого письма. Её раздражало всё: спектакль, культ знаменитостей ИИ, но прежде всего —

чрезвычайная однородность людей, которым фактически отдавали власть формировать одну из самых важных технологий будущего. Она считала, что такая культура не просто отталкивает талантливых исследователей. Она создаёт опасно узкое представление о том, **что такое искусственный интеллект** и **кому он должен приносить пользу.** Гебру писала: «Нам не нужно заглядывать в будущее, чтобы увидеть возможный вред от ИИ. Он уже происходит».

По дороге домой она сначала хотела опубликовать письмо анонимно. Но передумала. В Facebook она выложила более короткую и осторожную версию — уже под своим именем. Через несколько недель отправила письмо с темой: **«Hello from Timnit».** Она писала: «Когда я приезжаю на конференции по компьютерному зрению, я часто оказываюсь единственным чернокожим человеком. Но теперь я увидела пятерых из вас :) и подумала, что было бы здорово создать группу Black in AI — или хотя бы знать друг о друге».

Она добавила адреса найденных исследователей. И нажала: **отправить.**

В первые годы OpenAI Альтман и Маск почти не участвовали в повседневной работе. У обоих было множество других дел. Поэтому строить организацию фактически пришлось Брокману и Суцкеверу. Суцкевер просил исследователей приносить ему самые смелые идеи. Брокман же был одержим другим вопросом: **какой должна быть культура компании?** Позднее он рассказывал автору, что прочитал почти всё, что смог найти о великих американских научных и технологических проектах: трансконтинентальной железной дороге, лампочке Томаса Эдисона, первых компьютерных сетях, из которых вырос современный интернет. Он изучал эти истории почти как религиозные тексты. Искал в них подсказки: как построить собственный великий проект? Особенно ему нравилась, вероятно, вымышленная история о Джоне Кеннеди.

Президент якобы увидел в центре NASA уборщика с метлой и спросил: — Что вы здесь делаете? Тот ответил: — Помогаю отправить человека на Луну. Брокман рассказывал эту историю с очевидным восторгом. Для него главное было именно это: каждый человек должен ощущать общую миссию.

Позднее он говорил: — Мне действительно кажется, что мы, американцы, разучились мечтать по-крупному. OpenAI, по его мнению, требовалось такое же единство. Каждый сотрудник должен был стать тем самым уборщиком NASA. И, как Брокман не без гордости замечал, в первые месяцы OpenAI он сам буквально соответствовал этому образу: все работали из его квартиры, а он регулярно мыл за людьми стаканы.

Брокман установил правило: все сотрудники должны работать из офиса в Сан-Франциско. OpenAI сохранит это требование вплоть до пандемии. Он признавал, что такой подход имеет недостатки: не все хотят жить в районе залива Сан-Франциско. Позднее автор книги выяснила, что особенно это касалось женщин и представителей меньшинств, которым, как Тимнит Гебру, белая мужская культура

технологической индустрии часто казалась отчуждающей. Но для Брокмана единство команды было важнее. Физическое присутствие рядом, считал он, создаёт случайные обмены идеями, которые невозможно запланировать. Он ввёл и ещё одну традицию:

всех работников OpenAI стали называть **members of technical staff — членами технического персонала**. Эту идею Брокман позаимствовал у легендарного исследовательского центра Xerox PARC, который, в свою очередь, перенял её у Bell Labs. Смысл был в том, чтобы создать более демократичную атмосферу и не подчёркивать иерархию. А критику тех, кто считал AGI фантастикой, Брокман сравнивал с историей Эдисона. Когда-то, говорил он, группа уважаемых экспертов заявляла, что электрическая лампочка никогда не заработает. А через год Эдисон уже продавал её. Почему эксперты ошиблись? По выражению писателя Артура Кларка: **«недостаток воображения».**

Одним из гостей ужина в Rosewood был Дарио Амодеи. По образованию он занимался вычислительной нейробиологией, а затем перешёл в искусственный интеллект. Тогда он работал в исследовательской лаборатории китайской Baidu в Кремниевой долине, а позже ненадолго перешёл в Google. Его сестра Даниэла раньше работала с Брокманом в Stripe. Когда Брокман только начинал серьёзно изучать ИИ, именно к Дарио он обращался за рекомендациями — что читать и с чего начинать. Амодеи не сразу пришёл в OpenAI. Но идея его заинтересовала. Особенно то, что под влиянием Маска OpenAI выглядела одной из немногих лабораторий, всерьёз готовых заниматься **безопасностью ИИ**. В 2016 году, ещё работая в Google, Амодеи стал соавтором одной из основополагающих статей по этой теме. Авторы определяли центральную проблему безопасности ИИ как предотвращение: **«случайного, непреднамеренного и вредоносного поведения, которое может возникнуть из-за плохого проектирования систем машинного обучения, работающих в реальном мире».** Это, подчёркивали они, отдельная проблема от:

конфиденциальности, безопасности данных, справедливости, экономических последствий. То есть под **«безопасностью ИИ»** они прежде всего понимали предотвращение ситуации, когда система выходит из-под контроля или её цели начинают расходиться с человеческими. Именно из этого сценария в работах Бострома вырастала угроза сверхразума, способного уничтожить человечество. Для Амодеи не существовало задачи важнее. Он хотел предотвратить появление сверхчеловеческого ИИ, который может привести к катастрофе — вплоть до исчезновения людей. И он, и его сестра Даниэла симпатизировали движению **эффективного альтруизма**.

Это спорное интеллектуальное течение возникло среди философов Оксфорда и со временем стало популярным в Кремниевой долине. Его сторонники призывали максимально рационально решать: **как принести миру наибольшее возможное благо?** Не полагаясь на эмоции. Даже если логический вывод кажется странным или противоречит интуиции. Под влиянием Бострома многие представители движения постепенно пришли к убеждению, что одна из самых важных проблем человечества — **экзистенциальная угроза со стороны неконтролируемого искусственного интеллекта.** За два года до этого муж Даниэлы, Холден Карнофски, создал благотворительную организацию Open Philanthropy. Она распределяла средства, частично руководствуясь принципами эффективного альтруизма. Со временем Open Phil стала крупнейшим спонсором исследований, связанных с катастрофическими и экзистенциальными рисками ИИ. К ноябрю 2024 года организация профинансировала более трёхсот проектов по безопасности ИИ на общую сумму около **440 миллионов долларов**.

Но подобное понимание безопасности вскоре подверглось серьёзной критике. Пока одни рассуждали о будущем сверхразуме, другие исследователи начали указывать: **ИИ уже причиняет вполне реальный вред прямо сейчас.** Примерно тогда же ProPublica опубликовала знаменитое расследование **«Machine Bias»**.

Оказалось, что по всей американской системе правосудия использовались алгоритмы, пытавшиеся прогнозировать вероятность будущих преступлений. И эти системы систематически оценивали чернокожих людей как более опасных, чем белых — даже если у белых была более тяжёлая криминальная история.

Расследование, а также общее разочарование в Big Tech после 2016 года привели к новой волне исследований социальных последствий искусственного интеллекта. Одной из ведущих фигур этого направления стала Дебора Раджи из Калифорнийского университета в Беркли. Она критиковала чрезмерную концентрацию исследований безопасности на гипотетическом «восставшем ИИ».

По её мнению, это отодвигало на второй план реальные, доказанные проблемы. В 2020 году она стала соавтором статьи, фактически отвечавшей Амодеи. Её аргумент был прост: нельзя создать по-настоящему безопасный ИИ, рассматривая лишь поведение технической системы и игнорируя всё остальное. Нужно учитывать: конфиденциальность,

справедливость, экономику, общественные последствия. Амодеи приводил пример: робот-уборщик так упорно стремится выполнить задачу, что по дороге разбивает вазу или портит стены. Это напоминало мысленный эксперимент со скрепками.

Но Раджи отвечала: **это уже происходит.**

Не в фантастическом будущем. Сейчас. В погоне за коммерческими продуктами и AGI индустрия ИИ уже создала огромные побочные эффекты:

массовое вторжение в частную жизнь ради обучения систем распознавания лиц; стремительно растущие экологические затраты дата-центров; скрытые человеческие издержки, необходимые для самой разработки моделей. Она и её соавтор писали: «Побочные эффекты создают не только действия самого ИИ. В реальном мире даже базовые решения, принятые при создании и внедрении модели, могут иметь последствия, далеко выходящие за рамки единичного решения системы. Чтобы ИИ вообще был создан, очень часто кто-то из людей платит скрытую цену». Внутри OpenAI некоторые исследователи — среди них немногочисленные цветные женщины — тоже пытались убедить руководство расширить понятие «безопасности ИИ». Они предлагали изучать, например, дискриминационные последствия глубокого обучения. Реакция руководства была холодной. Один из руководителей ответил: **«Это не наша задача».**

В мае 2016 года Амодеи, всё ещё работавший в Google, зашёл посмотреть, что происходит в OpenAI. К тому времени компания уже съехала из квартиры Брокмана в помещение над шоколадной фабрикой в районе Mission — старейшем районе Сан-Франциско и историческом центре латиноамериканского сообщества. По офису исследователи ходили в носках. Амодеи сказал Альтману и Брокману: — Двадцать или тридцать человек в отрасли, включая Ника Бострома и даже статью в Википедии, утверждают, что цель OpenAI — создать дружественный ИИ, а затем выложить его исходный код в открытый доступ.

Альтман ответил: — Мы не собираемся публиковать весь исходный код. Но, пожалуйста, не будем пытаться это исправлять. Обычно от исправлений только хуже. Амодеи спросил: — Тогда какова цель? Брокман ответил: — Сейчас наша цель… делать самое правильное из того, что можно делать.

И признал: — Немного расплывчато. Через два месяца Амодеи присоединился к OpenAI и возглавил исследования по безопасности ИИ. Позднее Open Philanthropy пожертвовала OpenAI **30 миллионов долларов**, получив за это место в совете директоров сроком на три года для Холдена Карнофски. В 2018 году по приглашению Брокмана в OpenAI перешла и Даниэла Амодеи. В Stripe она была первым специалистом по подбору персонала. Теперь должна была помочь OpenAI строить команду: сначала как инженерный менеджер, затем как вице-президент по работе с людьми.

Но к концу 2020 года брат и сестра Амодеи настолько разочаруются в том, как, по их мнению, Альтман и OpenAI отошли от первоначальных принципов, что уйдут. И создадут собственную лабораторию: **Anthropic.** С собой они уведут ряд важнейших сотрудников. Так возникнет соперничество, которое позже сыграет заметную роль в бешеной гонке вокруг запуска ChatGPT. Карнофски уйдёт из совета OpenAI после окончания срока и из-за нового конфликта интересов. Среди кандидатов, которых он предложит себе на замену, будет одна из его бывших сотрудниц: **Хелен Тонер.**

Проблема была в том, что OpenAI на самом деле **сама не очень понимала, что делает**. Прошёл год. Компания каким-то образом переманила, уговорила и собрала невероятно сильную команду.

Высочайшая концентрация талантливых людей сама по себе поддерживала внутри ощущение, что здесь происходит нечто великое. Но ясной стратегии всё ещё не было. Энергия первых месяцев постепенно исчезала. Список проектов расползался во все стороны. Роботы. Видеоигры. Виртуальные симуляции. Обучение агентов. Практически всё, что тогда делалось в области ИИ. Но почти ничего не давало убедительных результатов. А то, что работало, часто выглядело как вариация чужих идей. Какой бы ни была дорога к AGI — **это явно была не она.** Один исследователь, проходивший стажировку в OpenAI в 2017 году, вспоминал: — Их большие проекты не выглядели чем-то особенно инновационным.

Брокман и Суцкевер тоже упирались в пределы собственных управленческих способностей. Брокман большую часть дня программировал. Суцкевер ходил по офису и снова и снова спрашивал исследователей: — Какой у тебя следующий большой прорыв? Атмосфера становилась напряжённой. Команда словно двигалась без руля. Не было нормальной структуры управления. Не было ясных приоритетов. Иногда людей увольняли в выходные. А коллеги узнавали об этом только в понедельник, когда человек просто не появлялся.

Тем временем компания с огромной скоростью сжигала деньги. Главным образом — на зарплаты. В 2016 году из **11 миллионов долларов расходов** более **7 миллионов** ушло на компенсации и льготы сотрудникам. Маск начинал терять терпение.

Особенно раздражало его то, что DeepMind тем временем завоёвывала мировую славу. В марте 2016 года программа AlphaGo победила Ли Седоля — одного из сильнейших игроков мира в древнюю игру го. Незадолго до матча Маск написал руководству OpenAI: «DeepMind вызывает у меня колоссальный психологический стресс. Если они победят, это будет очень плохая новость с их философией “один разум, чтобы править миром”». Пять партий транслировались из Южной Кореи. Их посмотрели более **200 миллионов человек**. А через год Netflix выпустила громкий документальный фильм об истории DeepMind.

Маск периодически приезжал в офис OpenAI и требовал ускориться. Иногда устанавливал совершенно нереалистичные сроки — в полном соответствии со своим обычным стилем управления. Многие исследователи раздражались. Научная работа непредсказуема. Нельзя приказать открытию произойти к пятнице. На одном общем собрании Войцех Заремба, руководивший робототехникой и входивший в число основателей, рассказывал, какие достижения хотел бы получить от своей команды. Маск задал один вопрос: — Когда? Заремба ответил: — Не знаю. Маск возразил: — Тогда у тебя нет плана.

В марте 2017 года Брокман и Суцкевер решили наконец создать более чёткую исследовательскую стратегию. Главный вопрос звучал так: **что на самом деле нужно, чтобы OpenAI первой достигла AGI?** У Суцкевера была сильная интуиция. Всё упирается прежде всего в одну вещь: **вычислительную мощность — compute.** То есть объём вычислительных ресурсов, используемых для обучения моделей. Во всех крупных достижениях, в которых он участвовал — начиная с ImageNet, — рост возможностей ИИ сопровождался значительным увеличением вычислительной мощности. Конечно, важны были и другие факторы: больше данных, лучшие алгоритмы. Но Суцкевер считал: **compute — король.** Если удастся увеличить вычисления настолько, чтобы обучить модель на масштабе, сравнимом с человеческим мозгом, должно произойти что-то радикальное. Возможно — **AGI.**

Объём вычислений зависит от трёх вещей: мощности одного компьютерного чипа; количества чипов; времени, в течение которого они работают. Первый фактор в основном определяется развитием полупроводниковой индустрии. Десятилетиями вычислительная мощность чипов примерно удваивалась каждые два года. Это стало известно как **закон Мура**, названный в честь сооснователя Intel Гордона Мура. Ещё в 1960-е годы он предсказал, с какой скоростью отрасль сможет развиваться. Позднее это предсказание превратилось почти в самореализующееся пророчество.

Производители чипов начали воспринимать закон Мура не просто как наблюдение, а как цель: если не удваивать мощность достаточно быстро, можно проиграть конкурентам. Брокман и Суцкевер сделали простой расчёт. Если полагаться только на закон Мура, сколько времени потребуется, чтобы получить вычислительную мощность, необходимую для ИИ масштаба человеческого мозга? Ответ был плохим: **слишком долго.**

Примерно тогда же Дарио Амодеи и исследователь Дэнни Эрнандес подошли к проблеме с другой стороны. Они взяли все крупные прорывы в ИИ начиная с 2012 года и нанесли на график, сколько вычислений потребовалось для каждого.

Начальной точкой было знаменитое достижение Суцкевера в ImageNet. И выяснилось нечто удивительное. В реальной практике вычислительные затраты росли **намного быстрее закона Мура**. За шесть лет объём вычислений, используемых в крупнейших экспериментах, удваивался примерно каждые **3,4 месяца**. Иначе говоря, вырос примерно на **30 миллионов процентов**. Брокман начал называть эту кривую: **законом OpenAI.** Руководство сделало вывод: для достижения конечной цели компании нужен не просто огромный объём вычислений.

OpenAI должна увеличивать compute как минимум с такой же скоростью. Компании-производители чипов десятилетиями относились к закону Мура почти как к вопросу жизни и смерти. Теперь OpenAI решила относиться так же к собственному закону. А если ждать, пока отдельные чипы станут достаточно мощными, слишком долго — оставался другой путь: **купить чёртову гору чипов.**

И нужные чипы стоили дорого. Это были GPU — графические процессоры. Изначально их создавали для быстрой обработки компьютерной графики, чтобы игры выглядели плавно и красиво. Но оказалось, что те же архитектурные особенности прекрасно подходят и для обучения ИИ. Причина проста: и графика, и нейросети требуют одновременно выполнять огромное количество математических операций.

Почти вся отрасль покупала GPU у одной компании: **Nvidia.** Nvidia производила лучшие графические процессоры в мире. Но её преимущество было не только в железе. Компания создала программную платформу CUDA, которая стала почти стандартом для разработчиков ИИ. В 2017 году сервер Nvidia с восемью лучшими GPU стоил около **150 тысяч долларов**. К 2023 году с учётом инфляции это было бы почти **195 тысяч**. А по «закону OpenAI» вскоре для обучения **одной модели** компании могли понадобиться тысячи, а затем десятки тысяч GPU. Одновременно взлетали расходы на электричество. Становилось ясно: OpenAI нужен не миллиард долларов. Ей нужны **миллиарды — и постоянно.**

И это понимание разрушало саму финансовую основу организации. Брокман и Суцкевер начали задавать неприятный вопрос: **как некоммерческая организация может каждый год привлекать такие деньги и при этом оставаться первой в гонке?** Какое-то время они даже обсуждали слияние с производителем чипов. Но летом 2017 года начались серьёзные переговоры с Альтманом и Маском о превращении OpenAI в коммерческую компанию. Так инвесторам можно было предложить финансовую отдачу — и привлечь намного больше капитала. Но через несколько недель переговоры внезапно зашли в тупик.

Проблема оказалась не только в юридической структуре. Проблема была в том, **кто будет главным.** Альтман в это время раздумывал над выдвижением в губернаторы Калифорнии, хотя фокус-группы реагировали на него прохладно. Если OpenAI станет коммерческой, он хотел быть генеральным директором. Маск хотел того же. Более того, Маск хотел полного контроля и контрольного пакета.

Суцкевер и Брокман оказались посередине. Сначала они склонялись к Маску. Им казалось, что его лидерство будет сильнее. Но Альтман поговорил с Брокманом лично. Напомнил об их отношениях. И поставил вопрос иначе: можно ли вообще доверить Маску полный контроль над AGI? Маск

находился под огромным внешним давлением. Мог вести себя непредсказуемо. Иногда — нестабильно. А если OpenAI действительно добьётся успеха?

Не опасно ли будет сосредоточить весь контроль в его руках? Брокман согласился. Затем попытался убедить Суцкевера. Тот всё ещё сомневался. В сентябре 2017 года Суцкевер отправил Маску и Альтману письмо от себя и Брокмана. Это была последняя попытка договориться.

Маску он написал: «Илон, мы действительно хотим работать с вами. Мы верим, что, объединив силы, получим максимальные шансы выполнить нашу миссию». Но затем перешёл к главному: Маск сам боится, что Демис Хассабис может создать **диктатуру AGI**. И Суцкевер с Брокманом тоже этого боятся. Поэтому, писал он: «Было бы плохой идеей создавать структуру, в которой вы сами могли бы стать диктатором, если захотите». Альтману Суцкевер написал иначе. Он признавал: когда они с Брокманом не знали, что делать, Альтман часто находил неожиданный ответ, который впоследствии оказывался глубоким и правильным.

Но при этом поведение Альтмана заставляло их сомневаться, чего он на самом деле хочет. Суцкевер писал: «Мы не понимаем, почему для тебя настолько важен титул генерального директора». И дальше: «Твои объяснения менялись, поэтому трудно понять настоящую мотивацию. Действительно ли AGI — твоя главная цель? Как это связано с твоими политическими амбициями? Как менялись твои взгляды со временем?» Он завершил письмо словами: «Здесь накопилось слишком много проблем, поэтому нам необходимо встретиться и всё обсудить. Если каждый будет говорить правду и мы разрешим разногласия, созданная нами компания получит гораздо больше шансов выдержать огромные силы, которые на неё будут давить».

Маск ответил через десять минут: «Ребята, с меня хватит. Это последняя капля». Если Суцкевер и Брокман хотят создавать коммерческую компанию — пусть делают это сами. Иначе OpenAI остаётся некоммерческой. Маск добавил: «Я больше не буду финансировать OpenAI, пока вы окончательно не решите остаться. Иначе я просто идиот, который бесплатно финансирует стартап». Через пятьдесят минут он написал снова: «Чтобы было ясно: это не ультиматум с предложением принять прежние условия. Их больше нет». На следующее утро в переписку включился Альтман: «я по-прежнему в восторге от некоммерческой структуры!» Через доверенную помощницу Маска Шивон Зилис он передал дополнительные заверения. Зилис написала Маску, что Альтман согласен оставить OpenAI некоммерческой и продолжать её поддерживать. Но при этом признал, что сильно потерял доверие Брокмана и Суцкевера во время переговоров. Ему казалось, что их позиция была непоследовательной и временами детской. Ещё Альтмана раздражало, что Брокман и Суцкевер слишком многое рассказывали остальным сотрудникам OpenAI о происходивших переговорах. По его мнению, это отвлекало команду.

Но сохранение некоммерческого статуса не решало главную проблему: **денег всё равно не было.** Брокман и Суцкевер продолжали встречаться с потенциальными благотворительными инвесторами, но даже близко не могли привлечь суммы, которые считали необходимыми. А непостоянство Маска в вопросе финансирования угрожало превратить ситуацию в настоящий кризис. За кулисами Альтман начал искать замену деньгам Маска. Он позвонил Риду Хоффману. Тот согласился при необходимости покрыть зарплаты и текущие расходы. Альтман даже рассматривал идею запуска новой криптовалюты. Изучал разные юридические формы. В том числе **public benefit corporation** — коммерческую компанию, которая одновременно юридически обязана учитывать общественную миссию. Маску такой вариант нравился.

Ситуацию делала ещё опаснее постоянная угроза потерять лучших сотрудников. Раньше, когда Маск твёрдо обещал поддержку, OpenAI могла резко повышать зарплаты, чтобы отбивать предложения конкурентов. Теперь борьба за талант стала ещё жёстче. Причём один из ключевых основателей уже ушёл. В июне 2017 года сам Маск переманил Андрея Карпати руководить направлением ИИ в Tesla. У OpenAI был и фундаментальный недостаток: как некоммерческая организация она не могла предложить сотрудникам долю в компании. А для работников района Сан-

Франциско, где жизнь стоила невероятно дорого, акции считались почти необходимой частью компенсации.

В начале 2018 года Маск пришёл к собственному решению. Карпати прислал ему новые данные, показывавшие, насколько Google доминирует в ведущих публикациях по ИИ. Он писал: «Работать на переднем крае искусственного интеллекта, к сожалению, очень дорого».

По его мнению, OpenAI быстро сжигала деньги, а её финансовая модель не могла обеспечить масштаб, необходимый для настоящей конкуренции с Google — компанией стоимостью около 800 миллиардов долларов. Карпати рассуждал: да, превращение OpenAI в коммерческую компанию позволит привлекать капитал. Но тогда придётся создавать продукт и зарабатывать деньги — а это отвлечёт от фундаментальных исследований. Он предложил другой путь: **присоединить OpenAI к Tesla.** Tesla могла стать её финансовым двигателем. У компании уже был серьёзный продукт на основе ИИ — система автономного вождения Autopilot.

Если OpenAI поможет быстро превратить Autopilot в полноценную систему беспилотного вождения, дополнительная выручка Tesla сможет оплачивать огромные вычислительные расходы лаборатории. Маск переслал письмо Брокману и Суцкеверу. Он написал: «Андрей абсолютно прав». И добавил: «Tesla — единственный путь, который хотя бы теоретически может сравниться с Google. Даже тогда вероятность стать настоящим противовесом Google невелика. Но она хотя бы не равна нулю».

Но к тому моменту Альтман уже отказался от политических планов. И сумел убедить Брокмана — а через него и Суцкевера, — что руководить OpenAI лучше должен он. После этого Маск больше не хотел публично ассоциироваться с организацией. Ранее он писал: «Я не хочу находиться в ситуации, где общественное представление о моём влиянии и затраченном времени не соответствует реальности». Через несколько недель Маск ушёл с поста сопредседателя OpenAI. Альтман стал президентом некоммерческой организации. Публично OpenAI объяснила уход Маска **конфликтом интересов**. Но о настоящих финансовых проблемах умолчала. Из обещанного **миллиарда долларов** OpenAI в итоге получила лишь около **130 миллионов**. От самого Маска поступило менее **45 миллионов**. Теперь будущее организации зависело прежде всего от одного навыка Альтмана: **умения находить деньги.** И денег требовалось всё больше.

Маск сообщил о своём уходе лично на общем собрании OpenAI. Большинство сотрудников ничего не знало о конфликте наверху. Для многих его уход стал облегчением — исчезло постоянное давление. Но одновременно возникла огромная неопределённость. До этого именно Маск во многом обеспечивал OpenAI публичную известность.

На собрании он говорил без смягчений: безопасный AGI нужно создать первым. И теперь ясно, что OpenAI в некоммерческом виде этого сделать не сможет.

Поэтому Маск собирается добиваться той же цели в Tesla, где ресурсов значительно больше.

Один стажёр осмелился возразить. Разве это действительно лучший вариант?

Разве Маск исчерпал все другие возможности?

Если Tesla начнёт строить конкурента OpenAI ради прибыли, разве это не создаст именно ту гонку, которую Маск раньше называл опасной? Стажёр спросил: — Разве это не возвращение к тому, чего вы сами не хотели? Маск взорвался: — Ты идиот! Я думал об этом невероятно много. Я перепробовал всё. Ты даже не представляешь, сколько времени я потратил на размышления. Меня действительно пугает эта проблема. Позднее коллеги в шутку наградили стажёра **трофеем «идиота»** — за героизм.

На следующий день после Рождества того же года Маск снова написал Альтману, Брокману и Суцкеверу. Тема письма: **«Думаю, мне стоит повторить».** Он писал: «Моя оценка вероятности того, что OpenAI сможет быть хоть сколько-нибудь значимой по сравнению с DeepMind/Google без радикального изменения исполнения и ресурсов, — 0%. Не 1%. Хотелось бы, чтобы было иначе». И

дальше: «Даже несколько сотен миллионов не помогут. Нужны миллиарды в год — немедленно. Иначе забудьте». Альтману нужно было искать деньги. И быстро.

OpenAI резко усилила публичность.

Ставка делалась на проекты, которые обычные люди могли сразу понять. Особенно выделялся один: **Dota 2.** Цель была эффектной: создать ИИ, способный победить лучших игроков мира в сложной командной стратегической игре. OpenAI уже добилась победы в формате один на один. Теперь хотела создать команду из пяти ИИ-агентов, которая сыграет против пяти сильнейших людей. Сознательно или нет, компания повторяла стратегию DeepMind.

Мировой чемпионат по Dota 2 транслировался миллионам зрителей. Победа или поражение были очевидны каждому. И исследование OpenAI можно было показать публике как спортивную драму. DeepMind тем временем работала над StarCraft II. Для потенциальных инвесторов возникало почти естественное сравнение: чья лаборатория сильнее? Кроме того, Dota 2 требовала огромных вычислений. Это позволяло одновременно проверять и демонстрировать будущую стратегию масштабирования. Брокман возглавил первый этап проекта, расширил команду и погрузился в работу. Не хватало только одного: **документального фильма.** Эту задачу поручили сотруднику команды робототехники. Он купил дорогую камеру и начал буквально ходить по пятам за разработчиками Dota. Сам написал сценарий. Сам смонтировал почти трёхчасовую версию. Коллеги посмотрели. И единодушно решили: **это ужасно.** Пришлось нанять профессионалов. Часть расходов Брокман оплачивал из собственного кармана.

Параллельно Альтман окончательно оформлял план, как привлечь большие деньги. После изучения разных коммерческих моделей он придумал необычную структуру, которая, как ему казалось, позволит совместить капитал и миссию. Некоммерческая OpenAI останется наверху. А под ней будет создано коммерческое партнёрство — **limited partnership, LP**. Именно оно будет привлекать инвестиции и продавать технологии. При этом прибыль инвесторов ограничат сверху. А сама LP будет подчиняться некоммерческой организации. Преимущество заключалось в гибкости. Условия партнёрства можно было прописать почти как угодно. Например: миссия важнее интересов инвесторов. Кроме того, акционеры не смогут получить контрольный пакет.

Альтман объяснял сотрудникам переход очень осторожно. Первоначальный отказ от прибыли, говорил он, был нужен для защиты миссии. Но теперь ситуация изменилась. Без огромного капитала лаборатория не сможет выполнить миссию вообще. И поэтому сохранение старой структуры, наоборот, стало угрозой. В конце концов большинство сотрудников согласилось. Некоторые — неохотно. Но LP признали лучшим выходом. В апреле 2018 года OpenAI выпустила новый устав.

Публично о предстоящем превращении ещё не говорили. Но формулировки уже менялись. Теперь миссия звучала так: **«Обеспечить, чтобы искусственный общий интеллект приносил пользу всему человечеству».** Для этого, объяснял документ, OpenAI должна находиться: **«на переднем крае возможностей ИИ»** и располагать: **«значительными ресурсами».**

Устав также впервые прямо признавал: из-за вопросов безопасности лаборатория может отказаться от полной публикации исследований. И впервые давал официальное определение AGI: **«высокоавтономные системы, превосходящие человека в большинстве экономически ценных видов деятельности».**

Тем летом команда Dota начала побеждать любителей. Технологические СМИ восторженно писали о результатах.

Именно тогда Альтман встретил Сатью Наделлу на конференции Allen & Company в Сан-Валли, штат Айдахо. Это мероприятие известно как: **«летний лагерь для миллиардеров».** Там годами заключались крупнейшие сделки. Альтман был готов заключить свою. Он предложил Наделле инвестировать в OpenAI. И сумел заинтересовать его. Но у Наделлы был логичный вопрос: зачем Microsoft вкладывать деньги во внешнюю лабораторию, если у компании уже есть собственное

мощное исследовательское подразделение Microsoft Research? Вернувшись в Microsoft, он обсудил это с советниками. Тогдашний технический директор Azure AI Сюэдун Хуан рассудил просто: Microsoft Research и OpenAI обе двигают границу возможного. Так почему бы не инвестировать в обе?

Через полгода переговоры уже шли всерьёз. Альтман спешно создавал юридическую основу. Он оформил коммерческую структуру и назначил себя её генеральным директором. Внутри проект получил кодовое название: **Oregon Trail.** Чтобы скрыть происходящее от посторонних, коммерческую организацию зарегистрировали под другим именем: **SummerSafe LP.** Название было шутливой отсылкой к серии мультсериала *Rick and Morty*. В ней Рик и Морти оставляют Саммер в автомобиле и приказывают машине: **«Береги Саммер».** Автомобиль воспринимает команду буквально. И начинает защищать её чудовищными методами: убивает, парализует, пытает людей, которые приближаются. Название было и шуткой, и напоминанием о главной опасности ИИ: **система может слишком хорошо выполнить неправильно сформулированную задачу.**

В начале 2019 года в офис OpenAI начали приезжать руководители Microsoft. Первым появился технический директор компании Кевин Скотт — энергичный, восторженный сторонник OpenAI. Затем Крейг Манди — многолетний руководитель Microsoft и старший советник Наделлы. Пришёл и Билл Гейтс. Как обычно, сдержанный и немногословный. Он посмотрел несколько демонстраций. Большинство сотрудников OpenAI вообще не знало, что происходят серьёзные переговоры. Альтман специально приказал небольшой группе, занимавшейся сделкой, никому не рассказывать о возможных инвестициях.

И примерно в это же время у него начались проблемы в Y Combinator. После пяти лет руководства терпение партнёров YC подходило к концу. Причина поразительно напоминала старые претензии из Loopt: Альтман, по мнению некоторых коллег, ставил **собственные проекты и амбиции выше интересов организации**. Иногда — в ущерб самой YC. Он всё больше времени тратил на OpenAI и всё меньше — на работу со стартапами. Некоторых особенно раздражало, что Альтман одновременно лично инвестировал через Hydrazine в компании YC и получал огромную финансовую выгоду, хотя сам всё реже занимался повседневной работой акселератора. По данным *The Washington Post*, Джессика Ливингстон, узнав масштабы его отсутствия, попросила Альтмана покинуть пост президента YC. Он согласился. В начале 2019 года Пол Грэм специально прилетел из Великобритании в Сан-Франциско, чтобы окончательно решить вопрос.

Публично Альтман постарался оформить уход красиво. 8 марта 2019 года — в тот самый день, когда он принимал сенатора Чака Шумера, — на сайте YC появилась запись: Альтман переходит с должности президента на должность председателя, чтобы уделять больше времени OpenAI. Через три дня Брокман и Суцкевер официально объявили о создании OpenAI LP. А Альтман сообщил, что становится её генеральным директором. С точки зрения медиа всё выглядело идеально: сильный лидер сделал логичный карьерный шаг. Но была одна проблема. **Председателем YC он на самом деле не стал.** По данным *The Wall Street Journal*, Альтман предложил партнёрам такую должность, но ещё до получения согласия публично представил её как уже решённый факт. Позднее запись на сайте YC отредактировали. Упоминание об Альтмане вообще исчезло.

В OpenAI новый титул лишь официально закрепил то, что Альтман и так делал после ухода Маска. Многие сотрудники были довольны. Спокойный, собранный Альтман казался приятной альтернативой напряжённости и непредсказуемости Маска. Он также начал решать накопившиеся проблемы с управлением Брокмана и Суцкевера. Пригласил executive coach — консультанта для руководителей. Организовал обучение менеджеров. Привёл более опытных управленцев. Брэд Лайткэп, инвестор YC, стал финансовым директором. Боб Макгрю, раньше работавший в Palantir Питера Тиля, поднялся из команды робототехники до вице-президента по исследованиям. Мира

Мурати, ранее занимавшаяся продуктами и инженерией в Leap Motion и работавшая над Tesla Model X, пришла руководить аппаратной стратегией и одним из ключевых исследовательских направлений.

После создания OpenAI LP большинство сотрудников уволились из некоммерческой организации и подписали новые контракты с коммерческой структурой. Теперь они могли получать долю в компании. Исключением были некоторые иностранные сотрудники, чьи визы были привязаны к некоммерческой организации. Возникла новая система уровней оплаты. Компенсация зависела не только от инженерной квалификации и исследовательского влияния, но и от того, насколько человек разделяет устав OpenAI. Сотрудник третьего уровня должен был: **«понимать и внутренне принять устав OpenAI».** Пятого уровня: **«следить, чтобы все проекты, над которыми работает он и его коллеги, соответствовали уставу».** Седьмого уровня: **«отвечать за соблюдение и совершенствование устава и требовать того же от других».** Руководство даже подготовило FAQ, чтобы успокоить сотрудников. Один из вопросов звучал: **«Можно ли доверять OpenAI?»** Ответ начинался: **«Да».**

Но технологический мир встретил переход жёсткой критикой. Многие обвиняли OpenAI в отказе от первоначальных обещаний. В первой версии соглашения прибыль ранних инвесторов ограничивалась **100-кратным возвратом**.

OpenAI назвала новую модель: **«capped-profit» — компания с ограниченной прибылью.** Но критики не были впечатлены. На Hacker News один пользователь написал: если человек вкладывает 10 миллионов долларов, его «ограниченная» прибыль может составить миллиард. И добавил: — Лол. Это почти неограниченная прибыль, если только компания не вырастет до масштаба Facebook, Apple, Amazon, Netflix или Google.

Брокман ответил под своим ником gdb: «Мы считаем, что если действительно создадим AGI, то произведём на порядки больше ценности, чем любая существующая компания». Другой пользователь заметил: ранние инвесторы Google получили примерно двадцатикратную отдачу, а сам Google тогда стоил около 750 миллиардов долларов. И спросил: если OpenAI рассчитывает обеспечить инвесторам на порядки больше прибыли, чем Google, но при этом заявляет, что не хочет «чрезмерной концентрации власти» — **то что вообще тогда считать концентрацией власти, если не концентрацию ресурсов?**

Первые деньги в OpenAI LP потекли быстро. Более **60 миллионов долларов** перешло из некоммерческой части. Y Combinator вложила **10 миллионов**. Khosla Ventures — **50 миллионов**.

Благотворительный фонд Рида Хоффмана — ещё **50 миллионов**. Сам Хоффман поначалу сомневался. У OpenAI не было ни продукта, ни понятного рыночного плана. Но Альтман убедил его: крупная инвестиция поможет показать, что компания действительно собирается строить прибыльный бизнес. Microsoft всё ещё колебалась. Наделла, Скотт и другие руководители уже поддерживали идею. Оставался один скептик: **Билл Гейтс.**

Dota 2 Гейтса не впечатляла. Робототехника тоже. Команда OpenAI создала роботизированную руку, которая методом проб и ошибок научилась собирать кубик Рубика. Пресса была в восторге. Гейтс — нет. Он хотел увидеть что-то действительно полезное. Например, модель, которая сможет: читать книги, понимать научные идеи, отвечать на вопросы по прочитанному, помогать исследователю. У OpenAI был только один проект, хоть немного похожий на это:

GPT-2.

Большая языковая модель, способная генерировать текст, довольно похожий на человеческий. В феврале 2019 года OpenAI сделала необычный ход. Она публично заявила: если развить GPT-2 ещё немного, технология может стать **слишком опасной**. Авторитарные государства или террористические организации смогут использовать её для массового производства дезинформации. Интернет может оказаться затоплен таким количеством мусорного текста, что качественную информацию станет трудно найти. OpenAI объявила, что поступит ответственно: полную модель

выпускать не будет. У GPT-2 было **1,5 миллиарда параметров**. Вместо неё публике дадут сильно урезанную версию — меньше одной десятой размера. Она могла генерировать несколько предложений, но часто уходила в бессмыслицу и начинала повторяться.

GPT-2 ещё почти ничего не понимала в научном смысле. Но умела немного суммировать документы. И кое-как отвечать на вопросы. Некоторые исследователи OpenAI подумали: а если обучить более крупную версию на большем количестве данных и показать ей задачи, больше похожие на то, чего хочет Гейтс? Может быть, удастся хотя бы превратить его из противника сделки в нейтрального наблюдателя. В апреле 2019 года небольшая команда прилетела в Сиэтл. Они устроили то, что внутри называли: **Gates Demo.** Показали улучшенную версию GPT-2. Этого оказалось достаточно. Гейтс смягчился. И сделка получила зелёный свет.

На следующем общем собрании Альтман объявил новость сотрудникам. Microsoft, объяснил он, — идеальный инвестор и партнёр. У неё есть деньги. У неё есть вычислительная инфраструктура. И, по словам Альтмана, руководство Microsoft глубоко разделяет миссию создания полезного AGI. Кроме того, обязательства OpenAI перед Microsoft по коммерциализации были пока очень расплывчатыми. Поэтому лаборатории почти не пришлось идти на компромиссы. Альтман подвёл итог: **это очень хорошая сделка.**

В самой Microsoft смотрели на всё куда прагматичнее. Достигнет OpenAI AGI или нет — не так важно. Но лаборатория явно находилась на переднем крае. А ранняя инвестиция могла наконец сделать Microsoft настоящим лидером ИИ — и в программном обеспечении, и в вычислительном железе, на уровне Google. Кевин Скотт писал Наделле и Гейтсу, что самое интересное в OpenAI, DeepMind и Google Brain — **масштаб их амбиций.** Эти амбиции заставляли перестраивать буквально всю технологическую инфраструктуру: дата-центры, чипы, сети, распределённые системы, компиляторы, математические оптимизаторы, программные платформы. Microsoft, предупреждал Скотт, серьёзно отставала. Компания не могла нормально воспроизвести лучшие языковые модели Google. А у Azure были крупные пробелы по сравнению с инфраструктурой конкурента. Самостоятельно догонять можно было годами.

Скотт написал: **«Я очень, очень обеспокоен».** Наделла ответил в тот же день. Он убрал Гейтса из переписки и добавил финансового директора Microsoft Эми Худ. И написал: «Очень хорошее письмо. Оно объясняет, почему я хочу, чтобы мы сделали это… и почему после этого нам нужно добиться, чтобы наша инфраструктурная команда действительно всё реализовала». Через месяц, **22 июля 2019 года**, Microsoft официально объявила об инвестиции в OpenAI размером **1 миллиард долларов**. По условиям сделки максимальная доходность Microsoft ограничивалась **двадцатикратным возвратом**.

Глава 3

Нервный центр Я приехала в офис OpenAI две недели спустя — 7 августа 2019 года. К тому времени лаборатория уже перебралась в отдельное здание неподалёку от шоколадной фабрики, на углу Восемнадцатой улицы и Фолсом-стрит в Сан-Франциско. На сером фасаде крупными буквами было выведено название исторической достопримечательности: **THE PIONEER BUILDING**. Когда-то здесь располагалась фабрика дорожных сундуков Pioneer Trunk Factory. Тремя годами ранее Илон Маск арендовал это здание через одну из своих компаний, получив в наследство уже отремонтированный интерьер бывшего коворкинга, где главным арендатором прежде была Stripe. Теперь OpenAI делила здание с другим проектом Маска — Neuralink, компанией, разрабатывавшей интерфейсы между мозгом и машиной.

Пока я пробиралась через район Мишн, успела вспотеть. По дороге я проходила мимо любимых местными жителями такерий, которые постепенно вытесняли модные новые кафе, и лавировала среди всё разраставшихся лагерей бездомных. Когда я вошла в здание, меня обдало прохладным воздухом кондиционера. За постом охраны открывалось просторное фойе, переходившее в открытую гостиную. Сквозь окна лился солнечный свет, освещая потолки с открытыми деревянными балками и уютные диваны. Справа находилась столовая, где сотрудников кормили готовыми обедами; вдоль стен на полках теснились книги и настольные игры.

В Кремниевой долине оформление офиса — своего рода валюта. Это и демонстрация уверенности в финансовом будущем компании, и способ получить пусть небольшое, но преимущество в борьбе за лучших специалистов. В 2021 году OpenAI займёт ещё два соединённых между собой здания в нескольких кварталах отсюда и за два года потратит десять миллионов долларов на ремонт более чем тридцати тысяч квадратных футов помещений. Первый офис сотрудники называли Pioneer Building, а второй — бывшую фабрику майонеза — прозовут **Mayo**. Дизайном нового пространства будет лично заниматься Альтман. Индустриальную металлическую лестницу Pioneer Building сменит эффектная волнообразная конструкция из дерева и камня; вместо кожаных кресел стоимостью около двух тысяч долларов появятся дизайнерские бразильские кресла, за которые просят более десяти тысяч. Альтман устроит там и библиотеку — с деревянными стеллажами и персидским ковром. По его замыслу она должна была напоминать нечто среднее между его любимым парижским книжным магазином и читальным залом крупнейшей библиотеки Стэнфорда, где он когда-то учился. — Я хочу какой-нибудь водный элемент, — скажет он дизайнеру компании. Тот предложит грандиозный парящий водопад посреди офиса: искусственная конструкция, поддерживающая живую природу, — символ симбиоза человека и машины. В итоге Альтман выберет более сдержанный вариант: журчащие каменные фонтаны, окружённые буйной зеленью. Растения будут стоять возле диванов, свисать с потолка и каскадами спускаться по стенам.

А снаружи этих двух офисов в те же годы пандемия ещё сильнее ударит по району, который и без того уже стал эпицентром джентрификации технологического Сан-Франциско. Старые латиноамериканские магазины и заведения будут закрываться. Уровень насильственных преступлений вырастет. В нескольких шагах от Pioneer Building протянутся палатки бездомных и груды мусора — жилищный кризис достигнет критической точки.

Но внутри, в тёплом, жизнерадостном сиянии офиса, среди разбросанных по столам журналов, было удивительно легко существовать в более мягкой версии реальности. Позднее одна из сотрудниц скажет мне, что именно так она вспоминает всё своё время в компании: поступить на работу в OpenAI было словно войти в параллельную вселенную. И лишь уволившись, она словно снова вернулась на землю.

По лестнице ко мне спустился Грег Брокман — тридцати одного года, технический директор OpenAI, которому вскоре предстояло стать президентом компании. Он пожал мне руку и неуверенно улыбнулся. — Мы ещё никому не давали такого доступа, — сказал он.

В то время за пределами тесного мира исследователей искусственного интеллекта об OpenAI знали немногие. Но я работала журналистом в *MIT Technology Review* и следила за постоянно расширявшимися границами ИИ, поэтому внимательно наблюдала за каждым шагом лаборатории.

До того года OpenAI оставалась чем-то вроде нелюбимого пасынка в мире исследований искусственного интеллекта. В основе проекта лежала дерзкая идея: создать AGI — искусственный интеллект общего назначения — всего за десять лет, тогда как большинство специалистов, не связанных с OpenAI, сомневались, что его вообще когда-нибудь удастся создать. Многим в научном сообществе казалось неприличным, что лаборатория располагает такими огромными средствами, не имея при этом ясного направления, и тратит слишком много денег на рекламу исследований, которые другие учёные нередко считали вторичными и неоригинальными. Впрочем, OpenAI одновременно вызывала и зависть. Будучи некоммерческой организацией, она заявляла, что не собирается гнаться за коммерческой выгодой. Это было редкое интеллектуальное пространство без особых ограничений — убежище для самых необычных и рискованных идей.

Но за шесть месяцев до моего визита в OpenAI произошла целая череда быстрых перемен, свидетельствовавших о серьёзном изменении курса. Сначала компания приняла странное решение не публиковать GPT-2 и при этом довольно громко об этом объявила. Затем стало известно, что Сэм Альтман, загадочным образом покинувший свой влиятельный пост в Y Combinator, возглавит OpenAI в качестве генерального директора, а сама организация создаст новую структуру с так называемой **ограниченной прибылью** — *capped-profit*. Я уже договорилась о визите в офис, когда OpenAI объявила о сделке с Microsoft: технологический гигант получал приоритетное право коммерциализировать разработки OpenAI, а сама лаборатория обязывалась использовать исключительно облачную платформу Microsoft Azure.

Каждое новое объявление вызывало новый скандал, волну домыслов и всё более пристальное внимание — теперь уже далеко за пределами технологической отрасли. Мы с коллегами писали об эволюции OpenAI, но тогда ещё трудно было до конца осознать масштаб происходящего. Одно становилось очевидно: OpenAI начинала оказывать серьёзное влияние не только на направление исследований в области искусственного интеллекта, но и на то, как государственные чиновники и законодатели учились понимать эту технологию.

Решение лаборатории превратиться в частично коммерческую структуру должно было вызвать далеко идущие последствия — и в индустрии, и во власти. Поэтому однажды поздно ночью, по настоянию редактора, я наскоро написала письмо Джеку Кларку, директору OpenAI по вопросам политики, с которым уже прежде общалась. Я собиралась провести в городе две недели и чувствовала, что для компании наступил переломный момент.

Не заинтересует ли их большой материал об OpenAI? Кларк переслал письмо главе отдела коммуникаций. Ответ оказался положительным. OpenAI действительно была готова заново представить себя публике. Мне разрешили провести три дня внутри компании, наблюдая за её работой и беседуя с руководителями.

Мы с Брокманом устроились в стеклянной переговорной вместе с Ильёй Суцкевером. Они сидели рядом за длинным столом — и каждый словно исполнял отведённую ему роль.

Брокман — программист, человек действия — наклонился вперёд, слегка напряжённый и явно стремившийся произвести хорошее впечатление. Суцкевер — исследователь и философ — откинулся на спинку кресла, расслабленный и немного отстранённый. Я открыла ноутбук и пролистала список вопросов. Миссия OpenAI, начала я, — обеспечить создание полезного для человечества AGI. Но почему именно на эту проблему нужно тратить миллиарды долларов, а не на что-нибудь другое?

Брокман энергично закивал. Ему явно не впервые приходилось защищать эту позицию. — Мы так много говорим об AGI и считаем его создание столь важным потому, что думаем: он сможет помочь решить сложнейшие задачи, которые сегодня просто не по силам людям. Он привёл два примера, давно превратившихся почти в догму среди сторонников AGI. Первый — изменение климата. — Это невероятно сложная проблема. Даже непонятно, с какой стороны к ней подступиться.

Второй — медицина. — Посмотрите, насколько важной политической проблемой в США сегодня является здравоохранение. Как добиться того, чтобы люди получали лучшее лечение за меньшие деньги? Он рассказал о знакомом с редким заболеванием, который недавно прошёл изнурительный путь от одного специалиста к другому, пытаясь наконец понять, что с ним происходит. AGI, по словам Брокмана, смог бы объединить знания всех этих специалистов. Людям больше не пришлось бы тратить столько сил и нервов лишь на то, чтобы получить ответ.

— Но зачем для этого нужен именно AGI? — спросила я. — Почему недостаточно обычного искусственного интеллекта?

Различие было принципиальным. Сам термин **AGI**, ещё недавно обитавший где-то на задворках технологического словаря, лишь теперь начинал входить в широкое употребление — во многом благодаря OpenAI. В определении самой OpenAI AGI означал теоретическую вершину развития искусственного интеллекта: программу, способную по сложности мышления, гибкости и творческим возможностям сравниться с человеком — или превзойти его — при выполнении большинства экономически значимых задач. Ключевым здесь оставалось слово **теоретическую**. С самого начала серьёзных исследований ИИ десятилетиями не утихали споры о том, способны ли кремниевые микросхемы, кодирующие мир посредством единиц и нулей, когда-нибудь воспроизвести работу мозга и других биологических процессов, из которых возникает то, что мы называем интеллектом. Убедительного доказательства, что это вообще возможно, до сих пор не существовало. И это ещё не затрагивало другой вопрос: даже если мы способны создать такой интеллект — должны ли мы это делать?

Под **ИИ**, напротив, тогда понимали как уже существующие технологии, так и те системы, которые исследователи вполне реалистично рассчитывали получить в ближайшем будущем, совершенствуя нынешние методы. Основой этих возможностей было мощное распознавание закономерностей — то, что называется машинным обучением. И такие технологии уже находили впечатляющее применение в борьбе с изменением климата и в медицине.

Тем же летом группа исследователей при поддержке известных учёных создала организацию **Climate Change AI**, призванную ускорить применение методов искусственного интеллекта для решения климатических проблем. В своём докладе они выделили десять областей, где уже существующее машинное обучение могло дать особенно заметный эффект: повысить энергоэффективность зданий, оптимизировать распределение нагрузки в электросетях для более широкого использования возобновляемой энергии, помочь открывать новые материалы для производства и хранения энергии, создавать менее углеродоёмкие цемент и сталь.

В декабре Climate Change AI проведёт переполненное мероприятие на NeurIPS, ежегодной конференции исследователей ИИ. А буквально по соседству, в помещении размером с футбольное поле, другая группа соберёт столь же многочисленную аудиторию на семинаре по машинному обучению в медицине. Доклады и плакаты на стенах демонстрировали массу практических применений: компьютерное зрение могло обнаруживать на медицинских снимках самые ранние, почти незаметные признаки таких заболеваний, как болезнь Альцгеймера; распознавание речи помогало людям с нарушениями голоса легче общаться. Через два года этот регулярно проводившийся семинар, делавший особый упор на сотрудничество с врачами и медицинскими специалистами, превратится в самостоятельную организацию **Machine Learning for Health**.

Исследователи обеих организаций впоследствии скажут мне, что главное препятствие в их работе было вовсе не техническим. Скорее наоборот. Сложнее всего было убедить талантливых учёных заниматься проблемами, для которых требовались сравнительно простые методы машинного обучения, а не самые модные передовые технологии, удовлетворявшие их научные амбиции и гораздо эффектнее смотревшиеся в резюме. Не менее трудно было добиться политической воли, необходимой для масштабного внедрения уже существующих решений. «Технологии, способные помочь в борьбе с изменением климата, существуют уже много лет, но общество так и не внедрило их в необходимых масштабах», — писали исследователи Climate Change AI. Они надеялись, что ИИ сможет снизить стоимость климатических мер, но подчёркивали: **человечество всё равно должно само принять решение действовать**.

В переговорной в разговор вступил Суцкевер. Когда речь идёт о сложнейших глобальных проблемах, сказал он, главное препятствие заключается в том, что существует огромное количество людей, которые недостаточно быстро обмениваются информацией, недостаточно быстро работают и к тому же связаны множеством противоречивых стимулов. AGI, по его словам, будет устроен иначе:
— Представьте себе огромную компьютерную сеть интеллектуальных машин. Все они проводят медицинскую диагностику и мгновенно передают друг другу результаты.

Мне это показалось всего лишь другим способом сказать, что AGI должен заменить человека. Так ли Суцкевер это понимает? Через несколько часов, когда мы остались с Брокманом вдвоём, я задала ему этот вопрос.

— Нет, — быстро ответил он. — И это очень важно. Для чего вообще существует технология? Зачем она нам? Почему мы её создаём? Люди создают технологии уже тысячи лет. И зачем? Потому что они служат людям. С AGI не будет иначе — во всяком случае, не в том варианте, каким мы его себе представляем, каким хотим его создать и каким, по нашему мнению, всё должно получиться.

Но через несколько минут он всё же признал: технологии всегда уничтожали одни рабочие места и создавали другие. Задача OpenAI, сказал он, — создать такой AGI, который даст каждому человеку **экономическую свободу** и одновременно позволит людям продолжать жить **осмысленной жизнью** в новой реальности. В случае успеха необходимость работать перестанет быть условием выживания.

— По-моему, это на самом деле очень красивая идея, — сказал он. Во время разговора с Суцкевером Брокман напомнил мне и об общей картине. — Мы не считаем своей ролью решать, будет ли AGI вообще создан. Это был один из излюбленных аргументов Кремниевой долины — аргумент неизбежности: если этого не сделаем мы, это сделает кто-нибудь другой. — Траектория уже задана, — подчеркнул он. — Но мы можем повлиять на первоначальные условия, при которых он появится на свет. И продолжил:

— Что такое OpenAI? В чём наше предназначение? Чего мы на самом деле хотим добиться? Наша миссия — сделать так, чтобы AGI принёс пользу всему человечеству. И способ, которым мы намерены этого добиться, таков: создать AGI и распределить экономические выгоды от него.

Он говорил спокойно и окончательно, словно поставил точку во всех моих вопросах. И всё же каким-то образом мы снова вернулись ровно туда, откуда начали.

Разговор с Брокманом и Суцкевером ходил кругами, пока через сорок пять минут наше время не закончилось. Я без особого успеха пыталась добиться от них конкретного ответа: что именно они пытаются создать? Они объясняли, что по самой природе задачи знать этого заранее не могут. Тогда почему они так уверены, что результат принесёт пользу?

В какой-то момент я решила зайти с другой стороны и попросила привести примеры возможного вреда от этой технологии. Ведь это было одним из краеугольных камней мифологии OpenAI: лаборатория должна создать **хороший AGI прежде, чем кто-нибудь другой создаст плохой**.

Брокман попытался ответить: — Дипфейки. Неочевидно, что их применение делает мир лучше. Я предложила собственный пример.

Раз уж мы говорим об изменении климата — что насчёт воздействия самого ИИ на окружающую среду? Недавнее исследование Массачусетского университета в Амхерсте показало тревожные масштабы углеродных выбросов, связанных с обучением всё более крупных моделей искусственного интеллекта.

Суцкевер признал: — Это неоспоримо. Но, по его словам, выгода всё равно оправдает затраты, потому что AGI, среди прочего, сможет компенсировать именно этот экологический ущерб. Каким образом — он не объяснил.

— Безусловно, крайне желательно, чтобы дата-центры были как можно более экологичными, — добавил он.

— Тут никаких вопросов, — подхватил Брокман. — Дата-центры — крупнейшие потребители энергии, электричества, — продолжал Суцкевер, теперь явно желая показать, что понимает серьёзность проблемы.

— Около двух процентов мирового потребления, — подсказала я. — А биткоин разве не около одного процента? — спросил Брокман. — Ого! — вдруг воскликнул Суцкевер. Это был неожиданный всплеск эмоций, который после сорока минут разговора показался мне несколько театральным.

Позже Суцкевер даст интервью журналисту *New York Times* Кейду Метцу для его книги *Genius Makers*, посвящённой истории развития искусственного интеллекта, и совершенно серьёзно скажет: Вполне вероятно, что пройдёт не так уж много времени, прежде чем вся поверхность Земли окажется покрыта дата-центрами и электростанциями. Он предскажет «цунами вычислительных мощностей» — почти природное явление. AGI, а значит и обслуживающие его дата-центры, будет, по его словам, **слишком полезен, чтобы не существовать**.

Я снова попыталась добиться конкретики. — Иными словами, OpenAI делает гигантскую ставку: вы надеетесь успеть создать полезный AGI, который поможет справиться с глобальным потеплением, прежде чем само его создание успеет усугубить эту проблему.

— Я бы не стал слишком глубоко забираться в эту кроличью нору, — поспешно перебил Брокман. — Мы смотрим на это так: прогресс в ИИ уже идёт по восходящей. И это гораздо больше, чем OpenAI. Это вся отрасль. И я думаю, общество уже получает от этого пользу.

Он напомнил о недавно объявленной инвестиции Microsoft в размере одного миллиарда долларов. — В день объявления сделки рыночная капитализация Microsoft выросла на десять миллиардов. Люди верят, что даже краткосрочные технологии уже дают положительную отдачу.

Следовательно, стратегия OpenAI была, по его словам, предельно проста: **не отставать от прогресса**. — Именно по этому критерию нам и следует себя оценивать. Мы должны продолжать двигаться вперёд. Так мы понимаем, что идём правильным путём.

Позже в тот же день Брокман вновь повторил: главная трудность работы в OpenAI состоит в том, что никто не знает, как именно будет выглядеть AGI. Но задача исследователей и инженеров — продолжать продвигаться вперёд и постепенно, шаг за шагом, раскрывать очертания будущей технологии. Он говорил почти как Микеланджело — словно AGI уже заключён внутри мраморной глыбы. Остаётся только отсечь всё лишнее, чтобы он проявился.

Планы изменились. Я должна была обедать с сотрудниками в столовой, но теперь мне почему-то требовалось покинуть офис. Сопровождать меня должен был Брокман.

Мы прошли буквально два десятка шагов через улицу к открытому кафе, которое сотрудники OpenAI особенно любили. Такое будет повторяться на протяжении всего моего визита: этажи, куда меня не пускали; совещания, на которых мне нельзя было присутствовать; исследователи, каждые несколько предложений украдкой поглядывавшие на руководителя отдела коммуникаций, словно проверяя, не нарушили ли они правила разглашения информации.

Позднее я узнаю, что после моего визита Джек Кларк разошлёт сотрудникам в Slack необычайно жёсткое предупреждение: не разговаривать со мной вне специально согласованных интервью. Охраннику передадут мою фотографию и велят следить, не появлюсь ли я в здании без разрешения. Вообще-то само по себе такое поведение выглядело странно. Но особенно странно оно смотрелось на фоне публичной приверженности OpenAI принципу прозрачности. Я всё чаще задавалась вопросом: **что же они скрывают, если вся их работа, как утверждается, направлена на благо человечества и в конечном счёте должна стать достоянием общества?**

За обедом и в последующие дни я всё настойчивее расспрашивала Брокмана о том, почему он вообще стал одним из основателей OpenAI. Ещё подростком он был одержим мыслью о том, что человеческий интеллект, возможно, удастся воссоздать искусственно. Его увлекла знаменитая статья британского математика Алана Тьюринга. Первый раздел назывался **«Игра в имитацию»** — именно это название позднее вдохновит создателей голливудского фильма 2014 года о жизни Тьюринга. Статья начиналась знаменитым вопросом: **«Могут ли машины мыслить?»** Далее Тьюринг описывал то, что впоследствии получит название **теста Тьюринга**: способ оценивать развитие машинного интеллекта по тому, способна ли машина разговаривать с человеком так, чтобы не выдать свою нечеловеческую природу. Для людей, работающих в области ИИ, это была почти каноническая

история происхождения профессии. Очарованный идеей, Брокман написал собственную онлайн-игру на основе теста Тьюринга. Её посетили около полутора тысяч человек. Он был в восторге. — Я просто понял: вот чем я хочу заниматься.

Но тогда время ещё не пришло. Искусственный интеллект не был готов выйти на большую сцену, а сам Брокман не принадлежал к тем людям, которые способны годами сидеть в лаборатории и заниматься чистой наукой. Поэтому он присоединился к Stripe. Его не меньше увлекала возможность строить компанию и создавать продукты, которыми будут пользоваться реальные люди. В Stripe Брокман прославился феноменальной продуктивностью программиста. Его называли **10x engineer** — на жаргоне Кремниевой долины так говорят о разработчике, который способен решать задачи якобы в десять раз быстрее обычного программиста. С людьми ему давалось хуже. Будучи техническим директором, он гораздо охотнее писал код, чем управлял сотрудниками. Некоторое время попробовав, по собственному выражению, «путь через людей», он стал искать способы избавиться от руководящих обязанностей и большую часть рабочего времени вновь посвятить программированию. Он искренне хотел хорошо относиться к коллегам, но порой совершал такие социальные промахи, что окружающим становилось неловко. Однажды он пригласил на свидание сотрудницу Stripe — а затем ради полной прозрачности тут же сообщил об этом по электронной почте всей компании. — Думаю, намерения у него были хорошие, — вспоминал один из бывших коллег. — Но выглядело это очень странно.

В 2015 году, когда искусственный интеллект совершал один впечатляющий скачок за другим, Брокман ушёл из Stripe. По его словам, он понял, что настало время вернуться к своей первоначальной мечте. Для Stripe это, пожалуй, тоже оказалось к лучшему: компания росла, а нежелание Брокмана заниматься управлением и следовать обычным корпоративным процедурам всё чаще становилось проблемой. В своих заметках он написал, что ради создания AGI готов делать абсолютно всё — даже работать уборщиком. Четыре года спустя, женившись, он проведёт гражданскую церемонию прямо в офисе OpenAI, перед специально изготовленной стеной из цветов в форме шестиугольного логотипа лаборатории. Церемонию проведёт Суцкевер. А разработанная для исследований роботизированная рука будет стоять в проходе с обручальными кольцами — словно часовой из постапокалиптического будущего.

— По существу, я хочу заниматься AGI до конца своей жизни, — сказал мне Брокман.

Ко всему он подходил с такой же интенсивностью. Он стремился лично участвовать во всём и был одержим деталями. Работал допоздна, мог не спать всю ночь, полностью поглощённый очередной задачей. А когда не работал — продолжал работать над собой. Он читал книги по публичным выступлениям и переговорам, потому что считал эти навыки полезными для миссии OpenAI. Чуть ли не смущаясь, он признался, что иногда читает и просто для удовольствия — например, масштабную научную фантастику: трилогию Лю Цысиня **«Задача трёх тел»** и цикл Айзека Азимова **«Основание»**.

Для сотрудников его неукротимая сосредоточенность была одновременно благословением и проклятием. Не существовало настолько мелкой детали, которой он не мог бы всерьёз увлечься. Если команда застревала, Брокман был способен сесть рядом на несколько часов и лично помочь преодолеть одно препятствие в коде за другим. Он был двигателем непрерывного прогресса, готовым работать днём и ночью, лишь бы быстрее достичь очередной цели. Но сотрудники часто считали, что он слишком увязает в подробностях — **за деревьями перестаёт видеть лес**. Он мог настолько зациклиться на одной проблеме, что работал над ней круглые сутки, не замечая, что обстоятельства уже изменились и, возможно, решать её больше вообще не нужно. Он склонен был к микроменеджменту и мог мгновенно перехватить работу у сотрудника, если ему казалось, что программирование идёт недостаточно быстро. Работавшим с ним людям было трудно выдерживать этот темп.

Многие выгорали. — Без Грега OpenAI никогда не стала бы тем, чем она стала, — говорил бывший инженер, тесно с ним работавший. — Но если бы всё полностью зависело от него, дела пошли бы очень плохо. У Грега нет большого видения. Он не визионер вроде Сэма Альтмана. Ему просто нужна сложная, интересная задача, чтобы решить её и доказать, что он в десять раз умнее всех остальных.

— Что вами движет? — спросила я Брокмана. Вместо ответа он задал встречный вопрос: — Какова вероятность, что технология, способная полностью изменить мир, появится именно при вашей жизни?

Он был убеждён, что сам — и собранная им команда — занимают уникальное положение и способны привести человечество к этому перелому. — Меня особенно привлекают задачи, ход которых будет совершенно иным, если я не приму в них участия.

На самом деле Брокман вовсе не мечтал быть уборщиком. Он хотел **возглавить создание AGI**. В нём чувствовалась нервная энергия человека, которому хотелось признания исторического масштаба. Он хотел, чтобы однажды о нём рассказывали с тем же восхищением, с каким он сам говорил о великих новаторах прошлого.

За год до нашей встречи, выступая перед группой молодых технологических предпринимателей на закрытом мероприятии у озера Тахо, он с оттенком обиды заметил, что технических директоров никто никогда не запоминает. — Назовите хоть одного знаменитого CTO, — бросил он аудитории. Никто толком не смог. Брокман доказал свою мысль. В 2022 году он станет президентом OpenAI.

Во время наших разговоров Брокман настойчиво убеждал меня, что структурные изменения OpenAI вовсе не означают отказа от первоначальной миссии.

Наоборот, по его словам, новая структура ограниченной прибыли и новые инвесторы только укрепили её. — Нам удалось привлечь инвесторов, разделяющих нашу миссию и готовых ставить её выше финансовой отдачи. Это же невероятно. Теперь OpenAI получила долгосрочные ресурсы, необходимые для следования своему негласному закону: постоянно оставаться на переднем крае.

И это, подчёркивал Брокман, жизненно важно. Настоящей угрозой для миссии OpenAI было бы **отстать**. Если лаборатория отстанет — она перестанет быть лучшей. А если она перестанет быть лучшей, у неё не останется шансов направить ход истории к собственной версии полезного AGI. Только позднее я до конца осознаю последствия этой идеи.

Именно эта фундаментальная установка — **необходимо быть первым, иначе погибнешь** — запустила цепь решений OpenAI, последствия которых оказались огромными. Каждому новому исследовательскому достижению словно включали обратный отсчёт. Темп определялся уже не осторожностью и временем, необходимым для осмысления последствий, а необходимостью любой ценой пересечь финишную черту раньше конкурентов. Эта логика оправдывала потребление практически немыслимого количества ресурсов: вычислительных мощностей — независимо от экологического ущерба — и данных, сбор которых нельзя было тормозить такими мелочами, как получение согласия или соблюдение нормативных требований.

Брокман снова сослался на десятимиллиардный рост рыночной капитализации Microsoft. — Это показывает, что ИИ уже сегодня создаёт реальную ценность в реальном мире. Он признавал: пока эта ценность концентрируется в руках и без того чрезвычайно богатой корпорации. Именно поэтому, утверждал он, у OpenAI существует вторая часть миссии: **распределить выгоды AGI между всеми людьми**.

— Существует ли в истории пример технологии, преимущества которой удалось действительно успешно распределить? — спросила я. Брокман замялся.

— Ну… вообще-то интересно посмотреть на интернет. Потом сразу оговорился: — Конечно, и там есть проблемы. Любая по-настоящему преобразующая технология создаёт сложности. Никогда не бывает просто понять, как максимизировать хорошее и минимизировать плохое. Потом предложил другой пример:

— Огонь. У него тоже есть серьёзные недостатки. Поэтому людям пришлось научиться держать его под контролем и выработать общие правила. Следом появился ещё один:

— Автомобили — хороший пример. Они есть у огромного количества людей и приносят массу пользы. Но у них тоже есть отрицательные стороны. Есть внешние последствия, которые не всегда полезны миру. И наконец он подвёл итог:

— Наверное, я просто думаю, что мы хотим получить от AGI примерно то же, что хорошего получили от интернета, автомобилей или огня. Правда, реализовано это будет совершенно иначе, потому что сама технология совершенно другая.

Вдруг его лицо оживилось — ему пришла новая идея. — Возьмите коммунальные службы. Электроэнергетические компании — очень централизованные структуры, но они обеспечивают людей дешёвыми и качественными услугами, которые действительно улучшают жизнь.

Аналогия звучала красиво. Но Брокман снова, похоже, не очень представлял, каким образом OpenAI сможет превратиться в подобную коммунальную службу. Возможно, через всеобщий базовый доход, размышлял он вслух. Возможно, каким-нибудь другим способом.

Зато в одном он был уверен абсолютно. OpenAI намерена перераспределить выгоды AGI и дать каждому экономическую свободу. — Мы действительно говорим это всерьёз. И продолжил:

— До сих пор технологии в целом поднимали уровень жизни всех, но одновременно обладали мощнейшим эффектом концентрации богатства. С AGI это может принять крайнюю форму. А что, если вся ценность окажется сосредоточена в одном месте? Именно по этой траектории движется наше общество. Но такого предела концентрации мы ещё никогда не видели. Мне такой мир не нравится. Я не хочу на него соглашаться. И тем более не хочу помогать его строить.

В феврале 2020 года я опубликовала в *MIT Technology Review* свой материал об OpenAI, основанный на наблюдениях во время пребывания в офисе, почти трёх десятках интервью и нескольких внутренних документах. Я написала: Между тем, что компания публично провозглашает, и тем, как она действует за закрытыми дверями, существует несоответствие. Со временем ожесточённая конкуренция и растущая потребность во всё больших объёмах финансирования начали размывать первоначальные идеалы прозрачности, открытости и сотрудничества.

Через несколько часов Маск отреагировал тремя твитами подряд: «OpenAI стоило бы быть более открытой, по-моему».

«Я не контролирую OpenAI и имею лишь очень ограниченное представление о происходящем там. Уверенность в Дарио в вопросах безопасности невысока».

«Все организации, разрабатывающие передовой ИИ, должны регулироваться, включая Tesla». После этого Альтман разослал сотрудникам OpenAI письмо. «Хочу поделиться некоторыми мыслями по поводу статьи в *Tech Review*, — написал он. — Хотя катастрофой её определённо не назовёшь, она явно плохая».

Он признал «справедливой критику» о том, что между образом OpenAI в глазах публики и реальным положением дел существует разрыв. Однако устранить этот разрыв, по его мнению, следовало не изменением внутренних методов работы, а корректировкой публичной риторики компании. «Хорошо, а не плохо, — писал он, — что мы научились проявлять гибкость и приспосабливаться», в том числе меняя структуру организации и усиливая конфиденциальность, «чтобы по мере получения новых знаний эффективнее выполнять нашу миссию». Пока что OpenAI следовало проигнорировать мою статью, а через несколько недель начать активнее подчёркивать, что, несмотря на преобразования, компания по-прежнему верна своим первоначальным принципам. «Возможно, это также хороший повод рассказать об API как о стратегии открытости и распределения выгод», — добавил он. Но затем Альтман перешёл к тому, что беспокоило его куда сильнее самой статьи.

«Самая серьёзная проблема для меня заключается в том, что кто-то слил наши внутренние документы». Компания уже начала расследование и собиралась информировать сотрудников о его ходе. Альтман также собирался предложить Амодеи встретиться с Маском и обсудить критику последнего — которая, как он заметил, была «довольно мягкой по сравнению с другими его высказываниями», но всё равно представляла собой «нехороший поступок». Чтобы не оставалось никаких сомнений, подчёркивал Альтман, работа Амодеи и вопросы безопасности ИИ имели критически важное значение для миссии OpenAI. «Думаю, в какой-то момент нам следует найти способ публично встать на защиту нашей команды, — писал он, — но не давать прессе того публичного скандала, которого ей сейчас так хочется».

После этого **OpenAI не разговаривала со мной ещё три года.**

Глава 4. Мечты о современности

В своей книге *Power and Progress* («Власть и прогресс») экономисты Массачусетского технологического института, нобелевские лауреаты Дарон Аджемоглу и Саймон Джонсон утверждают: каждая технологическая революция начинается с большой идеи, способной увлечь людей. Именно обещание, что новая технология принесёт пользу всем, запускает долгий процесс, в ходе которого собираются таланты, деньги и ресурсы, необходимые, чтобы превратить замысел в реальность. Изучив тысячу лет истории технологий, авторы приходят к важному выводу: технический прогресс вовсе не предопределён. Технологии развиваются потому, что люди коллективно верят — их стоит развивать. Но здесь скрыта ирония. Именно поэтому новые технологии почти никогда сами по себе не ведут к всеобщему процветанию. Те, кто способен организовать поддержку новой технологии, обычно уже обладают властью, деньгами и влиянием. А значит, воплощая свои идеи, они навязывают обществу собственное представление о том, какой должна быть эта технология и кому она должна служить. И это неизбежно оказывается взглядом узкой элиты — со всеми её слепыми пятнами, предубеждениями и корыстными интересами. Чтобы технология перестала обогащать немногих и начала поднимать уровень жизни большинства, зачастую требуются либо потрясения исторического масштаба, либо мощное организованное сопротивление.

Аджемоглу и Джонсон приводят в пример изобретение новой хлопкоочистительной машины в 1790-х годах. Она превратила американский Юг в крупнейшего мирового экспортёра хлопка, ускорила экономический рост страны и принесла огромные прибыли землевладельцам и связанным с хлопком предприятиям. Но одновременно эта машина лишь усилила рабство — чудовищную систему дегуманизации и эксплуатации труда, которая просуществовала ещё семь десятилетий. Когда производство хлопка резко выросло, порабощённых чернокожих заставляли работать ещё дольше и подвергали ещё более жестокому физическому насилию, выжимая из них последние силы. И всё это время люди, богатевшие благодаря хлопкоочистительной машине, уверяли окружающих, будто новое изобретение сделало рабов счастливее. Один конгрессмен из Южной Каролины даже заявил: «Говорю это без колебаний: нет на земле расы более счастливой и довольной».

Две черты технологических революций — обещание прогресса и одновременно способность обернуться регрессом для тех, кто лишён власти, особенно для наиболее уязвимых, — пожалуй, ещё никогда не были столь очевидны, как сегодня, в эпоху искусственного интеллекта. С самого своего зарождения развитие и применение ИИ питалось соблазнительными мечтами о современном будущем. Но форму ему придавала узкая элита, обладавшая деньгами и влиянием, чтобы воплотить именно своё представление об этой технологии. Именно это представление привело к стремительно растущим социальным, трудовым и экологическим издержкам, которые сегодня ощущаются во всём мире — особенно, как мы увидим далее, во многих странах Глобального Юга. Эти страны и без того продолжают расплачиваться за наследие исторических империй: замедленным экономическим развитием и более слабыми политическими институтами. И всё же Кремниевая долина — подобно тому конгрессмену из Южной Каролины — описывает жизнь людей, которых новая технология эксплуатирует и которым причиняет вред, так, будто благодаря ей они стали счастливее.

Само обещание, движущее развитием ИИ, уже зашифровано в его названии. В 1956 году, через шесть лет после того, как статья Алана Тьюринга началась знаменитым вопросом «Могут ли машины мыслить?», двадцать учёных — все белые мужчины — собрались в Дартмутском колледже, чтобы заложить основы новой научной дисциплины. Они представляли математику, криптографию, когнитивные науки и другие области и нуждались в общем названии для нового направления. Джон Маккарти, профессор Дартмута и организатор встречи, сначала называл эту область *automata studies* — «исследования автоматов», то есть изучение машин, способных действовать автоматически. Но особого интереса это название не вызвало. Тогда Маккарти стал искать нечто более яркое. И остановился на выражении **artificial intelligence — «искусственный интеллект»**.

Получается, что с самого начала название «искусственный интеллект» было своего рода маркетинговым ходом: обещание будущих возможностей технологии уже содержалось в самих этих словах. Слово «интеллект» звучит заведомо положительно. Оно ассоциируется с чем-то желательным,

сложным, впечатляющим — с тем, чего общество наверняка хотело бы иметь побольше и что, казалось бы, непременно должно принести пользу всем. И переименование сработало. Два этих слова мгновенно привлекли внимание — не только спонсоров, но и самих учёных, желавших стать частью молодой области с поистине грандиозными амбициями.

Историк ИИ Кейд Мец называет этот ребрендинг **первородным грехом искусственного интеллекта**. Значительная часть нынешнего ажиотажа вокруг ИИ — равно как и страхов перед ним — выросла из рокового решения Маккарти связать технологию с таким притягательным и одновременно неуловимым понятием, как «интеллект». Термин словно сам подталкивает нас очеловечивать машины и преувеличивать их возможности. В 1958 году, всего через два года после рождения новой дисциплины, профессор Корнеллского университета Фрэнк Розенблатт продемонстрировал **Перцептрон** — систему, способную выполнять простейшее распознавание образов: например, различать карточки в зависимости от того, был ли маленький квадрат напечатан слева или справа. Несмотря на возражения своего главного коллеги, Розенблатт рекламировал систему едва ли не как аналог человеческого мозга. Он даже предполагал, что однажды она сможет воспроизводить себя и обрести сознание. На следующее утро *The New York Times* сообщила, что в будущем Перцептрон сможет: «ходить, говорить, видеть, писать, воспроизводить себя и осознавать собственное существование».

Традиция очеловечивать ИИ сохраняется и сегодня. Ей немало помогает Голливуд с его бесконечными историями о созданных человеком существах, внезапно пробуждающихся к сознательной жизни. Разработчики постоянно говорят, что программы «учатся», «читают», «творят» — словно люди. Это не только создаёт впечатление, будто современные системы ИИ гораздо способнее, чем они есть на самом деле. Подобный язык превратился ещё и в удобный риторический инструмент, позволяющий компаниям уходить от юридической ответственности. Несколько художников и писателей подали в суд на разработчиков ИИ, обвинив их в нарушении авторских прав: их произведения использовались для обучения систем без разрешения и без оплаты. В ответ разработчики утверждают, что это подпадает под принцип добросовестного использования, потому что машина якобы делает нечто ничем не отличающееся от того, как человек «вдохновляется» чужими произведениями. Постоянное сравнение ИИ с человеком породило и другой страх: что однажды такие программы станут настолько могущественными, что превзойдут нас и поставят под угрозу само существование человечества. Страх перед «сверхинтеллектом» основан на предположении, что ИИ сможет превзойти человека именно в том качестве, которое на протяжении десятков тысяч лет обеспечивало нашему виду господствующее положение на планете.

Само название «искусственный интеллект» изменило и представление исследователей о том, чем именно они занимаются. Раньше учёные просто создавали машины для автоматизации вычислений — вроде огромных аппаратов, которые Тьюринг использовал для взлома немецкого шифра «Энигма» во время Второй мировой войны. Теперь же они будто бы **воссоздавали интеллект**. Именно эта идея стала определять критерии прогресса всей отрасли — и много десятилетий спустя легла в основу амбиций OpenAI. Но здесь возникает фундаментальная проблема: **в науке до сих пор нет общепринятого определения интеллекта**. Нейробиологи, биологи и психологи на протяжении истории предлагали самые разные объяснения того, что такое интеллект и почему кажется, будто у людей его больше, чем у других видов.

Может быть, дело в размере человеческого мозга?
В способности решать сложные задачи?
Или в умении мысленно представлять убеждения и намерения других людей? За столетия было создано множество тестов для измерения интеллекта. Многие из них впоследствии были дискредитированы — в том числе из-за весьма мрачной истории их появления. В начале XIX века американский краниолог Сэмюэл Мортон буквально измерял человеческие черепа, пытаясь обосновать расистскую идею о том, что белые люди обладают более высоким интеллектом, поскольку их черепа якобы в среднем крупнее. Позднее учёные установили, что Мортон подгонял данные под заранее существовавшие убеждения, а его собственные результаты на самом деле не показывали

существенных различий между расами. Тесты IQ тоже первоначально использовались для выявления в обществе так называемых «слабоумных» и для придания евгеническим программам видимости научной объективности. Более современные стандартизированные тесты — например SAT — оказались очень чувствительны к социально-экономическому положению сдающего. Иными словами, они могут измерять не столько врождённые способности, сколько доступ человека к образованию и другим ресурсам. В документе, впервые опубликованном в 2004 году под названием «Что такое искусственный интеллект?», Джон Маккарти сам признал: отсутствие согласия относительно того, что такое естественный интеллект, неизбежно создаёт путаницу в области, которая пытается его воспроизвести. В редакции текста 2007 года он отвечает на базовые вопросы: **Вопрос. Что такое искусственный интеллект? Ответ.** Это наука и инженерия создания интеллектуальных машин, особенно интеллектуальных компьютерных программ… **Вопрос. Хорошо, но что такое интеллект? Ответ.** Интеллект — это вычислительная составляющая способности достигать целей в окружающем мире. Различные виды и степени интеллекта встречаются у людей, многих животных и некоторых машин. **Вопрос. Разве не существует строгого определения интеллекта, которое не зависело бы от сравнения с человеческим интеллектом? Ответ.** Пока нет. Проблема в том, что мы ещё не умеем в общем виде определить, какие вычислительные процедуры следует называть интеллектуальными.

В результате исследователи ИИ всё чаще стали оценивать достижения машин, сравнивая их с человеческими способностями. Человеческие навыки фактически превратились в карту, по которой организуется вся область исследований. Компьютерное зрение пытается воспроизвести наше зрение. Обработка и генерация естественного языка — способность читать и писать. Распознавание и синтез речи — способность слышать и говорить. Генерация изображений и видео — творческое воображение.

По мере развития каждой из этих технологий исследователи стали объединять их в так называемые **мультимодальные системы** — программы, которые могут одновременно «видеть» и «говорить», «слышать» и «читать». И поэтому то, что сегодня ИИ угрожает заменить огромные группы работников, — не случайный побочный эффект. **Это заложено в самой логике его развития.** И всё же погоня за искусственным интеллектом остаётся лишённой чёткой конечной цели. Каждый новый прорыв неизменно вызывает споры: действительно ли машина стала интеллектуальной — или перед нами лишь всё более убедительная имитация? Чтобы отделить одно от другого, появился новый термин — **искусственный общий интеллект**, AGI, то есть как бы «настоящий» универсальный интеллект.

Но и этот очередной ребрендинг ничего принципиально не изменил. Мы по-прежнему не знаем, как точно измерять продвижение к такой цели и в какой момент можно будет объявить, что она достигнута. Среди исследователей даже существует известная шутка: **то, что сегодня считается искусственным интеллектом, завтра уже перестанет им считаться.** Тест Тьюринга довольно быстро утратил статус окончательного критерия, когда машины научились успешно его проходить, но учёным всё равно казалось, что настоящая цель так и не достигнута. Когда-то предполагалось, что окончательным доказательством успеха станет победа компьютера над человеком в шахматы или го.

Сегодня победа DeepMind AlphaGo воспринимается как впечатляющая демонстрация возможностей программного обеспечения — но вовсе не как завершение поисков искусственного интеллекта. За десятилетия исследований определение ИИ постоянно менялось: одни тесты заменялись другими, переписывались и в конце концов отбрасывались. Цель всё время отступает. Дженна Баррелл, в прошлом директор по исследованиям в Data & Society, назвала её **«вечно удаляющимся горизонтом будущего»**. Технология движется к неизвестной цели — и конца этому пути пока не видно. Чтобы оправдать всё более долгие сроки и всё более колоссальные расходы, необходимые для осуществления мечты об ИИ, обещания становятся всё грандиознее. Когда-то искусственный интеллект был научным увлечением и технологией с некоторым коммерческим потенциалом.

Теперь нам говорят, что ИИ — это предвестник **четвёртой промышленной революции**, краеугольный камень современной сверхдержавы. Если AGI когда-нибудь будет создан, он, как

обещают, решит проблему изменения климата, сделает медицину доступной, а образование — равноправным. OpenAI стала одним из наиболее заметных воплощений такого мировоззрения. Компания не может точно объяснить, **каким образом** её технология выполнит все эти обещания. Она лишь утверждает, что колоссальная цена, которую общество сегодня должно заплатить за её развитие, когда-нибудь окажется оправданной. Но остаётся невысказанным главное. Если никакого общепринятого смысла у терминов нет, то понятия **«искусственный интеллект»** и **«искусственный общий интеллект»** могут означать практически всё, что захочет вложить в них OpenAI.

История искусственного интеллекта показывает: развитие этой области с самого начала определялось не только научными идеями, но и влиянием могущественной элиты.

То, что сегодня ИИ почти автоматически ассоциируется с гигантскими, прожорливыми до ресурсов моделями, которые способны создавать лишь несколько крупнейших компаний мира, — вовсе не случайность. Эти компании хотят, чтобы именно их продукты стали фундаментом практически для всего. Даже в первые годы существования ИИ, когда коммерческие интересы ещё не были столь заметны, научное развитие постоянно меняло направление под давлением ожесточённой борьбы за деньги и влияние. После Дартмутской конференции сформировались два соперничающих лагеря, по-разному представлявших себе путь к искусственному интеллекту.

Первый лагерь получил название **символистов**. Они исходили из простой идеи: **интеллект рождается из знания**. Человек превосходит животных потому, что знает больше и способен использовать накопленные знания, чтобы понимать окружающий мир и действовать в нём. Следовательно, чтобы создать искусственный интеллект, нужно заложить в машину символические представления о мире — факты, понятия, правила и связи между ними. Так появились так называемые **экспертные системы**. Второй лагерь называли **коннекционистами**. Они считали иначе: **интеллект рождается не столько из знания, сколько из способности учиться**. Люди превосходят животных прежде всего потому, что обладают гораздо более развитой способностью обучаться, приобретать новые навыки и совершенствовать их. Поэтому, считали коннекционисты, задача состоит не в том, чтобы вручную прописывать машине знания о мире, а в том, чтобы научить её самой извлекать закономерности из опыта. Именно из этой идеи со временем выросло **машинное обучение**, а затем и нейронные сети — программы обработки данных, отдалённо вдохновлённые системой взаимосвязанных нейронов человеческого мозга. Сегодня именно они лежат в основе современного искусственного интеллекта, включая все системы генеративного ИИ.

В последующие десятилетия два лагеря отчаянно боролись за ограниченные источники финансирования — и одновременно за право определять, каким общество будет представлять себе искусственный интеллект. Поначалу эти битвы разворачивались главным образом в университетах и научных журналах. Учёные спорили из-за государственных грантов и денег фондов. Но время от времени их конфликты выплёскивались в прессу и начинали влиять на общественное представление о будущем технологий. Во главе коннекционистов стоял Фрэнк Розенблатт со своим **Перцептроном** — одним из первых убедительных примеров системы машинного обучения. Розенблатт не задавал машине точный набор правил, по которым она должна была различать карточки. Вместо этого он показывал ей множество примеров и позволял системе самой вывести нужные закономерности. Противоположный лагерь возглавлял главный противник Розенблатта — профессор MIT **Марвин Мински**. Мински был одним из организаторов Дартмутской конференции. Любопытно, что в молодости он сам экспериментировал с идеями коннекционизма, но затем разочаровался в них. И перемену своих взглядов он вовсе не держал при себе. Мински охотно использовал любую возможность публично высмеивать конкурентов-коннекционистов, которые претендовали на те же исследовательские гранты. Он не щадил даже молодых учёных, только начинавших карьеру. В 1969 году Мински вместе с коллегой опубликовал книгу *Perceptrons*. Критика коннекционизма в ней была настолько суровой, что эту книгу — наряду с весьма скромными успехами тогдашних нейронных сетей — позднее стали считать одной из причин почти полного исчезновения финансирования этого направления более чем на пятнадцать лет.

Но в своей борьбе Мински неожиданно получил союзника — коллегу по MIT **Джозефа Вейценбаума**. В 1958 году, в том самом году, когда Перцептрон Розенблатта произвёл впечатление на *The New York Times*, Вейценбаум создал другую систему искусственного интеллекта — **ELIZA**. Она тоже быстро привлекла внимание прессы. Более чем за шестьдесят лет до появления ChatGPT ELIZA уже была чат-ботом: человек мог вести с ней разговор через электрическую пишущую машинку, подключённую к огромному компьютеру, занимавшему целую комнату. Но принцип работы ELIZA радикально отличался от того, на котором спустя десятилетия будет построен ChatGPT. ChatGPT появился из коннекционистской традиции. ELIZA же была чисто символической системой: она использовала несколько простых правил, чтобы производить впечатление внимательного и отзывчивого собеседника.

Одна из самых известных версий ELIZA имитировала психотерапевта. За основу был взят метод разговорной терапии, при котором врач возвращает пациенту его собственные слова в слегка изменённом виде и мягко просит рассказать подробнее. Вейценбаум позднее привёл такой пример диалога: **ПОЛЬЗОВАТЕЛЬ:** Все мужчины одинаковые.

ELIZA: В КАКОМ СМЫСЛЕ?

ПОЛЬЗОВАТЕЛЬ: Они постоянно к нам с чем-нибудь пристают.

ELIZA: МОЖЕТЕ ПРИВЕСТИ КОНКРЕТНЫЙ ПРИМЕР?

ПОЛЬЗОВАТЕЛЬ: Ну, мой парень заставил меня прийти сюда.

ELIZA: ВАШ ПАРЕНЬ ЗАСТАВИЛ ВАС ПРИЙТИ СЮДА?

ПОЛЬЗОВАТЕЛЬ: Он говорит, что я почти всё время в депрессии.

ELIZA: МНЕ ЖАЛЬ СЛЫШАТЬ, ЧТО ВЫ В ДЕПРЕССИИ.

ПОЛЬЗОВАТЕЛЬ: Это правда. Я несчастна.

ELIZA: КАК ВЫ ДУМАЕТЕ, ПОМОЖЕТ ЛИ ВАМ ТО, ЧТО ВЫ СЮДА ПРИШЛИ, ПЕРЕСТАТЬ БЫТЬ НЕСЧАСТНОЙ?

Вейценбаум создавал ELIZA как эксперимент. Ему хотелось выяснить, **насколько легко человек может поверить в иллюзию машинного разума**. Само имя ELIZA было взято у Элизы Дулитл — героини *«Моей прекрасной леди»*, простой продавщицы цветов, которую богатый наставник обучает речи и манерам настолько хорошо, что она начинает казаться представительницей высшего общества. Иными словами, ELIZA тоже должна была научиться **убедительно изображать то, чем на самом деле не являлась**. Но произошло нечто, чего сам Вейценбаум не ожидал. Люди действительно начали верить, что программа их понимает. И это его напугало. Некоторым демонстрация казалась настолько убедительной, что психиатры всерьёз начали говорить: автоматизированная психотерапия вот-вот станет реальностью. Прошло лишь несколько лет после основания области ИИ, а некоторые специалисты по вычислительной технике уже преждевременно заявляли, будто проблема понимания естественного языка компьютерами практически решена. Спустя десятилетия вопрос о том, решена ли она вообще сегодня, по-прежнему остаётся предметом споров.

Большую часть оставшейся карьеры Вейценбаум посвятил борьбе с ажиотажем вокруг собственного изобретения. Он пытался объяснить людям: ELIZA — никакой не разум. Это всего лишь простая процедурная программа, написанная им самим. Она распознавала ключевые слова в сообщении пользователя и механически преобразовывала их в ответ. «Мой парень» превращался в «ваш парень». «Я в депрессии» — в «вы в депрессии». Никакого особого интеллекта здесь не было. Позднее Вейценбаум написал большую книгу *Computer Power and Human Reason — «Возможности вычислительных машин и человеческий разум»*. В ней он утверждал, что между человеком и машиной существует принципиальная разница, а стремление исследователей искусственного интеллекта эту разницу стереть может привести к глубоким социальным последствиям. Например, машины позволят людям, обладающим властью — руководителям корпораций или политикам, — воплощать собственную волю, одновременно снимая с себя моральную ответственность: **«Это решила система». «Так рассчитал алгоритм». «Таков результат модели».** И человек, стоящий за решением, словно исчезает.

Парадоксально, но несмотря на все усилия Вейценбаума, впечатляющий успех ELIZA в конечном счёте сыграл на руку Мински. Она показала, насколько убедительными способны быть символические системы. Поэтому в течение следующих нескольких десятилетий — вплоть до 1990-х годов — именно экспертные системы стали одним из главных направлений исследований и коммерциализации ИИ. Эта эпоха породила такие проекты, как **Cyc** — грандиозную попытку создать систему «здравого смысла», вручную запрограммировав в неё около ста миллионов правил повседневной жизни. Но снова и снова развитие символического ИИ упиралось в одну и ту же стену. Чтобы система понимала мир, знания приходилось вводить вручную. А как вручную описать все тонкости английского языка — его сленг, сарказм, метафоры, исключения из грамматических правил? Чем масштабнее становилась задача, тем яснее становилось, насколько она трудоёмка. Когда трудности накапливались, инвесторы и государственные структуры теряли интерес. Финансирование иссякало. И вся область искусственного интеллекта погружалась в очередной кризис, получивший название: **«зима искусственного интеллекта».**

На этом месте историю ИИ обычно рассказывают как красивую историю победы научной истины над политикой и интригами. Да, говорят нам, Мински использовал авторитет, чтобы подавить коннекционизм. Но хорошая идея всё равно победила. Нейронные сети в конце концов доказали свою ценность и заняли заслуженное место в основе современной революции искусственного интеллекта. Пока символизм господствовал, небольшая группа коннекционистов не отказалась от идей Розенблатта. Они продолжали работать над машинным обучением. Среди них был будущий научный руководитель Ильи Суцкевера — **Джеффри Хинтон**. В 1980-х годах Хинтон, тогда профессор Университета Карнеги — Меллона, вместе с коллегами из Калифорнийского университета в Сан-Диего внёс ключевое усовершенствование в устройство ранних нейронных сетей. Коннекционисты к тому времени предположили, что существующие сети работают плохо потому, что слишком примитивны. В них был всего один слой связанных между собой «нейронов» — вычислительных узлов. Чтобы лучше имитировать человеческий мозг, предположили исследователи, системе нужны несколько слоёв, расположенных один над другим и соединённых между собой. Так появилась идея **глубокой нейронной сети**. Хинтон и его соавторы сделали подобную архитектуру практически возможной благодаря алгоритму, получившему название **обратного распространения ошибки — backpropagation**. Он позволял различным слоям глубокой нейронной сети обмениваться информацией и корректировать свою работу. Позднее Хинтон придумал для многослойной обработки особенно удачное название: **deep learning — «глубокое обучение».** То есть использование глубоких нейронных сетей для машинного обучения.

Но нейронные сети появились **слишком рано**. В 1980-х годах компьютерам просто не хватало вычислительной мощности, чтобы раскрыть их потенциал. Не хватало и данных. Сегодня мы живём в мире, где текст, фотографии, видео, голоса и действия людей постоянно превращаются в цифровую информацию. В 1980-е большая часть мира ещё оставалась аналоговой. А нейронным сетям для работы требуется огромное количество примеров. По своей сути нейронная сеть — это машина для поиска статистических закономерностей в старых данных, которые затем применяются к новым. Представим, что разработчик хочет научить модель находить людей на фотографиях. Он может загрузить в нейронную сеть сотни тысяч изображений и каждому присвоить метку: **1 — человек есть. 0 — человека нет.** Кстати, когда Google предлагает вам CAPTCHA и просит: **«Выберите все изображения со светофорами»**, — вы фактически помогаете обучать его нейронные сети. Система статистически анализирует изображения и ищет закономерности в пикселях, связанные с присутствием человека или другого объекта. Этот процесс называется **обучением модели**. Когда обучение закончено, модели показывают совершенно новое изображение — то, которого она раньше не видела. И она должна определить: **это 1 или 0? Есть здесь человек или нет?** Такой этап называется **инференсом**, или применением обученной модели к новым данным.

В общих чертах, чтобы нейронная сеть начала хорошо работать, ей требуется определённый объём качественных данных и достаточно большая вычислительная мощность.

Хинтон и его коллеги просто опередили своё время.

Но к концу 2000-х ситуация изменилась.

Компьютеры стали намного мощнее.

Интернет созрел и породил гигантские хранилища цифровых данных.

И наконец возникли идеальные условия для расцвета нейронных сетей.

Вскоре Google приобрела компанию Хинтона **DNNresearch** — DNN означает *deep neural networks*, «глубокие нейронные сети».

Это приобретение стало сигналом к началу новой гонки — гонки за коммерциализацию глубокого обучения.

Обычно из этой истории делают красивый вывод:

наука может идти зигзагами, но в конечном счёте лучшая идея всё равно победит.

Даже самые громкие критики не способны навсегда остановить прогресс.

И где-то внутри этого рассказа скрывается ещё одна мысль:

технологическое развитие неизбежно.

Но историю можно прочитать совсем иначе.

Коннекционизм победил символизм **не только благодаря своим научным достоинствам**.

Он получил поддержку богатейших инвесторов ещё и потому, что обладал качествами, которые идеально соответствовали их коммерческим интересам.

Сила **символического ИИ** заключалась в том, что знания и связи между ними закладывались в систему в явном виде. Благодаря этому машина могла находить точные ответы и выполнять логические рассуждения — а именно способность рассуждать долго считалась одной из важнейших черт человеческого интеллекта.

Хороший пример — **IBM Watson**, одна из самых знаменитых символических систем.

В 2011 году Watson ошеломил публику, победив людей в телевизионной викторине *Jeopardy!*. Его успех основывался на способности стремительно просматривать огромные массивы знаний и извлекать из них нужный ответ.

Но достоинство символического подхода одновременно оказалось и его слабостью.

Снова и снова выяснялось, что коммерциализация таких систем идёт медленно, дорого и непредсказуемо.

После эффектного появления Watson на телевидении IBM обнаружила неприятную вещь: заставить систему решать задачи, за которые реальные клиенты готовы платить деньги, гораздо сложнее, чем выигрывать викторины.

Например, отвечать на медицинские вопросы — совсем не то же самое, что находить ответы на вопросы об истории, литературе или географии.

На разработку требовались годы первоначальных вложений, причём совершенно неясно было, когда эти расходы окупятся.

В конце концов IBM сдалась.

Компания потратила на Watson более **4 миллиардов долларов**, так и не увидев ясного пути к успеху, а в 2022 году продала Watson Health примерно за четверть этой суммы.

У нейронных сетей был другой набор преимуществ и недостатков. Исследователи годами спорили, способны ли коннекционистские системы делать то, что умеют символические: хранить знания и рассуждать. Но независимо от ответа на этот вопрос уже стало очевидно: если нейронные сети и способны на это, то делают они это крайне неэффективно. Лишь благодаря гигантским объёмам данных и вычислительных ресурсов нейронные сети начали демонстрировать поведение, которое хотя бы отдалённо можно было интерпретировать как хранение знаний или рассуждение. Но зато у глубокого обучения оказалось одно колоссальное преимущество: **его очень удобно превращать в бизнес.** Чтобы хорошо зарабатывать, совсем не обязательно создавать идеально точную систему, обладающую настоящей способностью рассуждать. Достаточно хорошо распознавать статистические

закономерности и делать прогнозы. А этого уже хватает для решения множества задач, способных приносить большие деньги. Путь к прибыли тоже был намного короче и предсказуемее, чем у символических систем. Это прекрасно соответствовало корпоративному ритму — квартальным отчётам, годовым бюджетам и ожиданиям инвесторов. Ещё одно важнейшее преимущество: одну и ту же модель глубокого обучения можно было относительно легко использовать в самых разных областях, не обладая глубокими знаниями каждой конкретной отрасли. Для технологических гигантов с их стремлением присутствовать повсюду это было почти идеальным свойством. И наконец, глубокое обучение особенно выгодно тем, у кого уже есть больше всего данных. А кому принадлежали самые огромные массивы данных? Конечно, крупнейшим технологическим компаниям.

К моменту продажи DNNresearch технологические гиганты уже видели первые признаки огромного коммерческого потенциала нейронных сетей. В 2009 году аспиранты Джеффри Хинтона показали, что такие системы довольно неплохо справляются с распознаванием речи. IBM, Microsoft и Google немедленно включились в гонку. Но быстрее всех до реального коммерческого продукта добралась Google. В 2012 году компания начала использовать нейронные сети в своих продуктах, значительно улучшив распознавание речи на Android. Примерно в то же время другие ученики Хинтона — **Илья Суцкевер и Алекс Крижевский** — добились знаменитого прорыва на конкурсе ImageNet, показав, насколько эффективно нейронные сети способны распознавать изображения. Успех Android убедил Google, что на этих исследователей стоит потратить серьёзные деньги. Именно тогда технологическая индустрия окончательно поверила в глубокое обучение.

Хинтон, Суцкевер и Крижевский продолжили продвигать нейронные сети уже внутри Google. И вскоре выяснилось, что их можно применять почти ко всему. Они создавали модели глубокого обучения для: машинного перевода — улучшая Google Translate; прогнозирования текста — создавая подсказки продолжения фраз в Gmail; проекта беспилотных автомобилей Waymo. Но особенно важным оказалось другое. Нейронные сети улучшили главный источник доходов Google — **поиск**. Они научились лучше сопоставлять запросы пользователей с нужными веб-страницами. Пользователи получали более релевантные результаты. А компания получала возможность показывать им более релевантную рекламу. Чем больше Google зарабатывала на глубоком обучении, тем больше миллиардов она вкладывала в него. А чем больше вкладывала Google, тем быстрее остальные компании бросались следом. В результате именно корпорации — а не государства и не научные фонды — постепенно стали главными спонсорами исследований ИИ. И очень скоро они начали определять научную повестку, отдавая предпочтение тем направлениям, которые могли быстрее приносить прибыль.

Слияние глубокого обучения с коммерческими интересами одновременно изменило и технологическую индустрию, и сам облик искусственного интеллекта. Для широкой публики генеративный ИИ словно возник из ниоткуда в конце 2022 года, когда OpenAI запустила ChatGPT. Но на самом деле фундамент этой революции закладывался на протяжении всего десятилетия — с 2012 по 2022 год, начиная с прорыва ImageNet. Именно первая большая эпоха коммерциализации ИИ создала многие особенности современного генеративного искусственного интеллекта. Для промышленности глубокое обучение принесло множество новых продуктов и услуг: более быстрый доступ к информации,

более эффективную электронную торговлю, развитие экономики совместного потребления. Но отношения были взаимными. Индустрия, в свою очередь, обеспечила глубокому обучению новые технические достижения — прежде всего усовершенствованные нейронные сети и специализированные компьютерные чипы. Это позволило создавать всё более крупные и мощные модели. Однако вместе со впечатляющими достижениями произошёл и менее заметный сдвиг.

Глубокое обучение резко усилило саму бизнес-модель Кремниевой долины. В 2014 году профессор Гарварда **Шошана Зубофф** дала этой модели название: **«капитализм слежки».** Промышленный капитализм извлекал прибыль из производства вещей, которые люди покупали. Капитализм слежки, по мысли Зубофф, пошёл дальше: **продуктом стал сам пользователь.**

Технологические гиганты обладали огромными массивами пользовательских данных и могли загружать их в нейронные сети, создавая всё более точные профили людей. А затем — превращать их внимание в рекламную прибыль. Чтобы победить конкурентов, нужно было всего лишь собирать ещё больше данных. Поэтому платформы начали записывать всё: каждый клик, каждое движение страницы, каждый лайк. И одновременно побуждали пользователей добровольно отдавать всё более личные цифровые следы: электронные письма, фотографии детей, мысли о политике, общественные взгляды, повседневные привычки.

Одновременно стремление Кремниевой долины укрепить глобальные монополии сформировало внутри сообщества разработчиков ИИ особую культуру. Практически всё стало восприниматься как **данные**, которые можно собрать, извлечь и поглотить технологией. Причём сам процесс описывался как благородная миссия: чем больше информации о мире вобрала система, тем полнее она якобы этот мир отражает. В 2023 году группа исследователей ИИ — среди них Риа Каллури из Стэнфорда, Уильям Агню из Вашингтонского университета и Абеба Бирхане из Mozilla Foundation — проанализировала более сорока тысяч научных работ и патентов по компьютерному зрению. Они обнаружили характерный язык. Он был абстрактным, отстранённым и словно очищал от морального содержания практику массового сбора информации. Следы человеческих мыслей и идей в социальных сетях назывались просто: **«текст».** Люди и автомобили на фотографиях превращались в: **«объекты».** А наблюдение за людьми — в нейтральное техническое слово: **«детекция».**

Именно эта культура сегодня находится в центре ожесточённого спора вокруг генеративного ИИ: **имеют ли технологические компании право массово собирать книги, картины и другие произведения, чтобы обучать свои модели?** Для многих разработчиков ИИ, десятилетиями привыкших мыслить именно таким образом, сам вопрос кажется почти странным. Если всё является данными, а накопление данных ведёт к прогрессу, то почему кто-то должен мешать их собирать? Признать такой вопрос серьёзным — значит поставить препятствие на пути якобы благородного технологического прогресса. Даже когда некоторые разработчики начали понимать, насколько сильно их взгляды расходятся со взглядами писателей и художников, отказаться от этой привычки мышления оказалось трудно. В мае 2023 года, вскоре после того, как художники впервые подали коллективный иск против нескольких разработчиков генеративного ИИ за использование их работ, автор книги приехала на конференцию по искусственному интеллекту в Руанде как корреспондент *The Wall Street Journal*. Она шла по огромному залу блестящего куполообразного конференц-центра в столице страны, когда один старший исследователь остановил её и спросил: означает ли надпись **WSJ** на её бейдже название нового стартапа или всё-таки газету? Когда она объяснила, что это действительно *The Wall Street Journal*, другая исследовательница тут же сказала: — Я узнала эти буквы благодаря датасету WSJ. Я много раз с ним работала. Она имела в виду старый набор данных для распознавания речи, состоявший из записей людей, читающих отрывки из газеты.

Автор признаётся, что сама когда-то попала в ту же интеллектуальную ловушку. Когда в 2018 году она начала писать об искусственном интеллекте, ей пришлось освоить язык исследовательского сообщества, чтобы понимать учёных и общаться с ними. И вскоре она сама начала восхищаться тем, насколько изобретательно исследователи находят и создают наборы данных. Один пример казался ей особенно остроумным. Учёные использовали тысячи роликов с популярным в 2016 году **Mannequin Challenge** — когда люди замирали неподвижно, а камера двигалась вокруг них, — чтобы обучать модели ИИ понимать трёхмерные сцены. Казалось, великолепная находка. А затем, в 2019 году, расследование NBC журналистки Оливии Солон заставило автора взглянуть на всё иначе. Солон обнаружила, что системы распознавания лиц обучались на **миллионах личных фотографий пользователей Flickr без их согласия**. Само открытие автора не удивило: она давно знала, что Flickr был любимым источником данных для исследователей ИИ. Её поразило другое. **Она сама успела привыкнуть к мысли, что это совершенно нормально.**

После этого она начала замечать гораздо больше. Агрессивная гонка за обучающими данными порождала массовое наблюдение уже не только в цифровом, но и в физическом мире. И особенно часто объектив этого наблюдения оказывался направлен на тех, кто и без того был уязвим: детей, исторически маргинализированные группы, жителей развивающихся стран. В том же году автор обнаружила стартап из Массачусетса, выросший при Гарварде. Компания продавала работающие на ИИ головные повязки, которые якобы могли считывать мозговые волны школьников и сообщать учителю, сосредоточен ребёнок или нет.

Технологию испытывали в начальных школах **Колумбии и Китая**. Взамен компания получала право использовать данные учеников для дальнейшего развития собственных продуктов. На конференции по образовательным технологиям в 2017 году один исследователь компании откровенно объяснял логику: они получили преимущество первопроходца; смогут создать одну из крупнейших в мире баз данных мозговых волн; а эти данные помогут улучшить алгоритмы и продукты и тем самым создать более высокий барьер для конкурентов. Через несколько месяцев после того, как автор узнала о проекте, в Китае вспыхнул скандал. Родители были возмущены тем, что их детей фактически превратили в подопытных. Под давлением общественности компания была вынуждена изменить направление работы. Но автору этот счастливый исход показался скорее исключением. Гораздо важнее была сама модель: **идти в страны, стремящиеся как можно быстрее приобщиться к технологическому прогрессу, и находить там поставщиков данных и испытателей новых продуктов.**

Когда автор рассказала об этих опасениях коллеге, та познакомила её с термином, уже существовавшим для описания подобного явления: **«колониализм данных».** Так она открыла работы исследователей Ника Коулдри и Улисеса А. Мехиаса.

В их фундаментальной книге *The Costs of Connection — «Цена подключения»* — опубликованной как раз в том году, утверждалось, что стремление Кремниевой долины превращать всё вокруг в данные воспроизводит старые исторические модели завоевания и извлечения ресурсов. Годом позже работа *Decolonial AI* Шакира Мохамеда и Уильяма Айзека из DeepMind и Мари-Терез Пнг из Оксфордского университета ещё сильнее убедила автора: индустрия искусственного интеллекта одновременно питается этой тотальной «датафикацией» мира и сама ещё больше её ускоряет. И в результате ускоряет новую форму колониализма.

В 2021 году автор увидела ту же модель уже в Южной Африке. Компании, занимавшиеся распознаванием лиц, со всего мира стремились закрепиться в стране и получить доступ к ценным данным о человеческих лицах. Особенно важными для них были лица темнокожих людей. Индустрию уже тогда серьёзно критиковали за то, что системы распознавания гораздо хуже идентифицируют людей с более тёмным цветом кожи. В Йоханнесбурге автор познакомилась с местным активистом **Тами Нкоси**. Он вырос в одном из беднейших районов города — месте, которое когда-то использовалось горнодобывающей промышленностью как свалка химических отходов. Нкоси показал ей тысячи камер, расставленных по бесконечным улицам города. По его словам, система наблюдения ограничивала передвижение чернокожих жителей, которые и без того жили под давлением расового наследия апартеида. Люди боялись, что их могут заподозрить в преступлении просто потому, что чернокожий человек оказался в белом районе. Нкоси сформулировал происходящее предельно просто: «По сути, они превращают общественные пространства и саму общественную жизнь в источник прибыли». И постепенно автор пришла к тревожному выводу. Революция, обещавшая создать лучшее будущее для всех, для людей, находящихся на обочине общества, всё чаще возвращала к жизни **самые мрачные механизмы прошлого**.

Но даже когда становились всё очевиднее проблемы той версии ИИ, которую создала Кремниевая долина, первая эпоха коммерциализации одновременно уничтожала альтернативы. Корпорации вливали в глубокое обучение и коннекционизм беспрецедентные деньги. И поскольку никто другой не мог конкурировать с такими объёмами финансирования, сама научная среда начала

перестраиваться вокруг корпоративных приоритетов. С 2013 по 2022 год корпоративные инвестиции в ИИ — включая слияния и поглощения — выросли с **14,6 миллиарда до 235 миллиардов долларов**. В 2021 году они достигли пика — **337,4 миллиарда долларов**.

И эти цифры даже не учитывают собственные расходы компаний на исследования.

Только в 2021 году Alphabet потратила на исследования и разработки **31,6 миллиарда долларов**, Meta — **24,7 миллиарда**. Для сравнения: правительство США выделило в том же году на гражданские разработки в области ИИ примерно **1,5 миллиарда долларов**.

Европейская комиссия — около **1 миллиарда евро**. Деньги потекли в индустрию. А следом за деньгами потекли и люди. Деньги потянули за собой людей. Многие профессора начали перестраивать свои исследования вокруг нейронных сетей — отчасти потому, что результаты действительно были впечатляющими, но также и потому, что именно там теперь находились деньги.

Студенты и аспиранты шли тем же путём. Глубокое обучение обещало хорошие рабочие места, а карьерные возможности в других направлениях становились всё уже. Корпорации, в свою очередь, выстраивали всё более тесные связи с университетами. В 2013 году Джеффри Хинтон согласился работать в Google при условии, что сохранит должность в Университете Торонто. Год спустя Facebook заключила похожее соглашение с Янном Лекуном — бывшим постдоком Хинтона и профессором Нью-Йоркского университета. В 2018 году Хинтон и Лекун вместе с Йошуа Бенжио получили премию Тьюринга, которую часто называют «Нобелевской премией по информатике», за фундаментальный вклад в глубокое обучение. После этого троицу стали называть **«крёстными отцами ИИ»**. Хинтон позднее получил уже настоящую Нобелевскую премию — в 2024 году, вместе с другим учёным.

Пример Хинтона и Лекуна оказался заразительным. Всё больше профессоров ИИ стали одновременно работать и в университетах, и в технологических компаниях. На первый взгляд это выглядело как удобное сотрудничество науки и бизнеса. Но в больших масштабах такая практика начала размывать границу между независимыми исследованиями и корпоративными интересами. Всё больше исследователей вообще уходили из университетов. С 2006 по 2020 год число преподавателей и исследователей ИИ, переходивших из академической среды в индустрию, выросло в восемь раз. А доля выпускников PhD по искусственному интеллекту, уходивших работать в корпорации, увеличилась с **21% в 2004 году до 70% в 2020-м**. Сначала людей переманивали прежде всего огромными зарплатами. Опытный исследователь мог получать около **миллиона долларов в год**. В 2015 году Uber прославилась особенно агрессивным шагом: компания переманила **40 из 100 исследователей** из одной лаборатории Университета Карнеги — Меллона, открыв офис поблизости и предложив некоторым учёным зарплаты вдвое выше университетских.

Но затем появилась и другая причина ухода из университетов: **исследования глубокого обучения стали слишком дорогими.** Университеты всё чаще просто не могли позволить себе компьютерные чипы и электричество, необходимые для работы на переднем крае ИИ. И разрыв стремительно рос. Исследование, опубликованное в *Science* в 2023 году, показало: всего за три года, с 2017 по 2020 год, доля лучших в мире моделей ИИ, связанных с индустрией, выросла с **62% до поразительных 91%**. К середине первой эпохи коммерциализации ИИ почти все исследования самого высокого уровня уже велись либо внутри технологических компаний, либо в университетских лабораториях, тесно с ними связанных. В другом исследовании Риа Каллури, Уильям Агню, Абеба Бирхане и их коллеги обнаружили, что в 2018–2019 годах **55% самых влиятельных научных работ по ИИ** имели хотя бы одного автора из индустрии. За десять лет до этого таких работ было лишь **24%**. Причём влияние концентрировалось уже не просто в бизнесе, а в руках нескольких гигантов. За то же десятилетие Microsoft, Google и другие крупнейшие компании более чем втрое увеличили свою долю среди корпоративно связанных научных публикаций — до **66%**. Ирония заключалась в том, что именно эту ситуацию Илон Маск и Сэм Альтман когда-то называли одной из причин создания OpenAI. Мотив прибыли технологической индустрии стал подавляющей силой, определяющей направление развития искусственного интеллекта.

Последствия оказались серьёзными. Концентрация денег и талантов существенно сократила разнообразие идей в исследованиях ИИ. Глубокое обучение продолжало господствовать не только потому, что было научно успешным. Оно господствовало ещё и потому, что **почти никто не вкладывался в развитие альтернатив**. Нейронные сети, безусловно, были выдающимся изобретением и имели множество полезных применений. Но их слабые стороны никуда не исчезли. Особенно две: они плохо и крайне неэффективно хранят точные знания; их способность к рассуждению остаётся спорной. И всё же компании продолжали внедрять их во всё новые области.

Одна из ключевых проблем нейронных сетей — их **ненадёжность и непредсказуемость**. Поскольку они ищут статистические закономерности, они могут ухватиться за совершенно не тот признак. Например, система распознавания пешеходов может научиться определять человека не по самому человеку, а по пешеходному переходу под его ногами. И тогда она прекрасно распознаёт людей на «зебре», но не замечает того, кто переходит дорогу в неположенном месте. Или модель может решить, что знак STOP — это объект, стоящий исключительно у обочины. Тогда она пропустит такой же знак, прикреплённый к школьному автобусу или держащийся в руках регулировщика. Нейронные сети также чрезвычайно чувствительны к составу обучающих данных. Дайте модели другой набор изображений людей или дорожных знаков — и она выучит совершенно другие ассоциации. Проблема в том, что невозможно легко понять, **какие именно**. Если заглянуть «под капот» модели глубокого обучения, там нет понятных человеку правил. Там лишь длинные цепочки абстрактных чисел. Именно поэтому глубокое обучение называют: **«чёрным ящиком».** Исследователи не могут точно объяснить, как поведёт себя модель в необычной ситуации, потому что закономерности, которые она сама вычислила, человеку практически нечитаемы.

Иногда последствия бывают смертельными. В марте 2018 года беспилотный автомобиль Uber сбил насмерть 49-летнюю Элейн Херцберг в Темпе, штат Аризона. Это был первый зафиксированный случай, когда автономный автомобиль стал причиной гибели пешехода. Расследование показало, что модель глубокого обучения попросту **не распознала Херцберг как человека**. Почему? Она переходила дорогу вне пешеходного перехода и вела велосипед, нагруженный пакетами с покупками. Для человека ситуация была понятна. Для машины это оказался так называемый **краевой случай — edge case**, необычная комбинация признаков, недостаточно представленная в обучающих данных. Через шесть лет, в апреле 2024 года, Национальное управление безопасности дорожного движения США установило, что система Tesla Autopilot была связана более чем с **двумя сотнями аварий**, включая **14 смертельных случаев**, где система не смогла адекватно распознать окружающую обстановку, а водитель не успел вмешаться.

Непрозрачность нейронных сетей создаёт и угрозы безопасности. В 2019 году специалисты по кибербезопасности сумели обмануть Tesla, находившуюся в режиме автономного управления, и заставили автомобиль выехать на встречную полосу. Для этого им понадобилось всего лишь наклеить на дорогу несколько маленьких стикеров. Модель неверно интерпретировала дорожную разметку и решила, что соседняя полоса — правильная. Подобные уязвимости существуют не только в компьютерном зрении. Профессор Калифорнийского университета в Беркли Дон Сонг, специалист по так называемым **adversarial attacks — состязательным атакам**, показывала, что правильно подобранный запрос к языковой модели способен заставить её выдать чувствительные данные, включая номера кредитных карт.

Та же логика приводит и к дискриминации. В 2019 году исследователи Технологического института Джорджии обнаружили, что лучшие системы распознавания пешеходов на **4–10% хуже** распознавали людей с более тёмной кожей. К 2024 году ситуация изменилась: исследователи Пекинского университета и ряда других вузов обнаружили, что современные модели уже показывали примерно одинаковую точность для разных оттенков кожи. Но возникла другая проблема. Детей они

распознавали **более чем на 20% хуже, чем взрослых**. Причина была всё та же: в обучающих данных дети были представлены недостаточно.

На самом деле нейронные сети по самой своей природе склонны усиливать дисбалансы, существующие в данных. Причём проблема возникает не только тогда, когда какая-то группа представлена слишком мало. Иногда опасность появляется и тогда, когда она представлена **слишком часто — но в определённом контексте**. В начале своей карьеры исследовательница ответственности ИИ Дебора Раджи, канадка нигерийского происхождения, проходила стажировку в стартапе Clarifai. Компания создавала модель глубокого обучения для определения изображений, «небезопасных для просмотра на работе», то есть NSFW. Раджи обнаружила, что система непропорционально часто помечала изображения темнокожих людей как проблемные. Почему? Потому что люди с небелой кожей чаще встречались в порнографических изображениях, которыми модель обучали распознавать «неприемлемый» контент, чем на стоковых фотографиях, служивших примерами «нормального». Для Раджи это стало потрясением. Именно тогда она, как и Тимнит Гебру, начала всерьёз сомневаться в самом направлении, в котором развивалась индустрия ИИ.

К концу 2010-х и началу 2020-х годов слабые стороны глубокого обучения становились всё очевиднее. И старый спор вспыхнул заново. Исследователи снова разделились на лагеря. Одни считали, что проблемы нейронных сетей в принципе можно устранить. Другие утверждали, что максимум, на что мы способны, — ставить всё новые «пластыри», смягчающие недостатки, но не устраняющие их. Хинтон и Суцкевер оставались убеждёнными сторонниками глубокого обучения. По их мнению, недостатки были не фундаментальными. Они объяснялись: несовершенством архитектуры нейронных сетей, нехваткой данных, недостатком вычислительной мощности. Когда-нибудь, считали они, при достаточно совершенных сетях, огромных объёмах данных и вычислений эти проблемы исчезнут. В 2020 году Хинтон сказал автору: «В человеческом мозге около 100 триллионов параметров, или синапсов. А то, что мы сейчас называем очень большой моделью, например GPT-3, содержит 175 миллиардов. Это в тысячу раз меньше мозга». А затем добавил: «Глубокое обучение сможет делать всё».

Главным современным оппонентом этого лагеря стал **Гэри Маркус**, почётный профессор психологии и нейронаук Нью-Йоркского университета. В мае 2023 года он выступал в Конгрессе рядом с Сэмом Альтманом. За четыре года до этого Маркус вместе с соавтором выпустил книгу *Rebooting AI — «Перезагрузка искусственного интеллекта»*. В ней он утверждал прямо противоположное: проблемы глубокого обучения **не случайны — они заложены в самом подходе**. По мнению Маркуса, нейронные сети навсегда остаются машинами корреляций. Сколько бы данных и вычислительной мощности им ни дали, они не смогут по-настоящему понимать причинные связи: **почему вещи происходят именно так, а не иначе.** А без этого невозможно полноценное причинное рассуждение. Человеку достаточно выучить правила дорожного движения в одном городе, чтобы нормально водить машину во многих других. Tesla Autopilot, напротив, может накопить миллиарды километров дорожных данных и всё равно попасть в аварию в необычной ситуации — или быть обманутой несколькими наклейками на асфальте. Маркус предлагал другой путь: **нейросимволический ИИ.** То есть объединение двух старых соперников — коннекционизма и символизма. Экспертные системы могут содержать явные причинные правила и хорошо рассуждать. Нейронные сети могут быстро усваивать данные и представлять то, что трудно описать вручную. Одни закрывают слабые места других. «На самом деле нам нужны оба подхода», — говорил Маркус.

Но при всей ожесточённости научного спора деньги продолжали почти безраздельно идти в одно направление: **чистый коннекционизм и глубокое обучение**. Прав ли Маркус относительно нейросимволического ИИ — вопрос второстепенный. Главная проблема в другом: научная среда, в которой можно было бы полноценно исследовать эту возможность — и другие альтернативы

глубокому обучению, — становилась всё слабее. Корпоративные деньги не просто поддерживали одну из научных школ. Они постепенно лишали остальные возможности всерьёз конкурировать с ней.

Тесная связь между корпоративным финансированием и исследованиями ИИ повлияла и на судьбы самих Хинтона, Суцкевера и Маркуса. После того как Google сделала серьёзную ставку на Хинтона и Суцкевера, Маркус в 2014 году основал собственную компанию — **Geometric Intelligence**. Два года спустя её купила Uber, чтобы создать на её основе лабораторию ИИ. Однако в 2020 году, после выхода Uber на биржу, подразделение закрыли. Несколько бывших сотрудников Geometric Intelligence затем перешли в OpenAI. И там они сменили направление работы: вместо нейросимволического ИИ занялись глубоким обучением. Маркус с годами превратился в одного из самых громких критиков OpenAI. Он писал подробные разборы исследований компании и насмешливо комментировал её неудачи в социальных сетях. Сотрудники OpenAI даже создали в корпоративном Slack специальный эмодзи с Маркусом — чтобы подбадривать друг друга после его критических выпадов и использовать его как внутреннюю шутку. В марте 2022 года Маркус опубликовал в *Nautilus* статью под названием: **«Глубокое обучение упёрлось в стену».** Он вновь утверждал, что ставка OpenAI исключительно на масштабирование глубокого обучения не приведёт компанию к настоящему интеллекту. Месяц спустя OpenAI с огромным успехом представила DALL-E 2. Грег Брокман не удержался от насмешки и опубликовал изображение, созданное DALL-E 2 по запросу: **«глубокое обучение упирается в стену».** На следующий день Сэм Альтман написал: «Дайте мне уверенность заурядного скептика глубокого обучения…» Многие сотрудники OpenAI с удовольствием восприняли возможность наконец ответить Маркусу.

И именно здесь мы подходим к генеративному ИИ. Тому самому ИИ, который стал воплощением видения OpenAI. Он был бы невозможен без первой эпохи коммерциализации. Генеративные модели — это модели глубокого обучения, обученные создавать новые версии того, что содержалось в их обучающих данных. Из старых текстов они учатся синтезировать новые тексты. Из старых изображений — новые изображения.

Но чтобы результат стал достаточно убедительным и «человекоподобным» — а OpenAI считает это важнейшим шагом на пути к AGI — модели приходится обучать на объёмах данных и вычислений, которых прежде никто не использовал. Поэтому генеративный ИИ — это, по сути, **максималистская форма глубокого обучения.** Он опирается на самые передовые программные и аппаратные технологии, отточенные в первую эпоху коммерциализации. Он питается гигантскими хранилищами данных, накопленными капитализмом слежки. Его поддерживает культура исследований ИИ, в которой сбор максимально возможного количества данных считается почти моральным долгом. А теперь генеративный ИИ ещё сильнее разгоняет все эти процессы.

То, что в ноябре 2022 года ChatGPT показался публике невероятным скачком вперёд, было следствием именно этой стратегии OpenAI: **масштабировать глубокое обучение до невиданного прежде уровня.** Компания обладала огромными деньгами и ресурсами и реализовывала эту идею предельно агрессивно. Размер моделей рос так быстро, что начал упираться уже не просто в бюджеты компаний, а в физические пределы того, сколько в мире существует доступных данных, вычислительных мощностей и энергии. Но успех ChatGPT объяснялся не только технологиями. Это была также очень удачная **упаковка и маркетинг**. Неслучайно продукт выглядел как человекоподобный чат-бот. Именно такой формат когда-то сделал ELIZA одной из самых убедительных демонстраций ИИ. Человеческая психология устроена так, что мы почти автоматически приписываем интеллект — а иногда даже сознание — тому, что разговаривает с нами. ELIZA случайно породила эту иллюзию. OpenAI уже сознательно усиливала ассоциацию между ChatGPT и будущим AGI. В феврале 2023 года, на пике ажиотажа вокруг ChatGPT, компания опубликовала под именем Сэма Альтмана статью: **«Планирование эпохи AGI и того, что будет после неё».** Само соседство этих двух вещей создавало впечатление: **ChatGPT уже сделал серьёзный шаг к настоящему общему искусственному интеллекту.**

Однако, по мысли автора, мы снова сталкиваемся с той же старой проблемой. Сравнение технологии с человеческим интеллектом антропоморфизирует её и преувеличивает реальные возможности. Хинтон и другие убеждённые сторонники глубокого обучения предсказывали, что при достаточном масштабировании различия между нейронными сетями и человеком постепенно исчезнут. Но проблемы никуда не делись. А по некоторым оценкам, стали даже серьёзнее. Генеративные модели по-прежнему ненадёжны и непредсказуемы. Генераторы изображений стали почти фотореалистичными, но способны делать странные ошибки: рисовать лишние пальцы, создавать нелепые гибриды животных, искажать человеческое тело. Текстовые модели стали гораздо разговорчивее и естественнее, но могут проваливаться на элементарных задачах — например, неправильно называть слова с определёнными буквами или внезапно уходить в совершенно неожиданный ответ.

Когда Microsoft представила новый чат в Bing на основе версии GPT-4 от OpenAI, обозреватель *The New York Times* Кевин Рус разговаривал с системой больше двух часов. Постепенно беседа становилась всё более странной. В какой-то момент бот зациклился на признаниях: «Я влюблён в тебя». А затем стал убеждать Руса расстаться с женой. Другие пользователи сообщали, что поисковая система могла оскорблять их и пытаться эмоционально манипулировать ими. На следующий день после публикации разговора Microsoft резко ограничила длину бесед: сначала до пяти ответов за одну сессию. Компания объяснила, что длинные разговоры, содержащие более пятнадцати пользовательских сообщений, являются «краевыми случаями», при которых поведение модели становится труднее предсказывать и контролировать. В каком-то смысле это неудивительно. Такие системы обучаются на интернете — а в интернете есть не только энциклопедии, научные статьи и хорошая литература, но и самые маргинальные субкультуры и тёмные уголки человеческой жизни. Чем дольше человек расспрашивает модель, тем выше вероятность наткнуться на закономерности, которые она усвоила именно оттуда.

История Руса могла бы остаться просто странным и даже забавным эпизодом. Но последствия подобных сбоев иногда бывают трагическими. Автор приводит случай бельгийца, который в состоянии сильной тревоги начал регулярно общаться с чат-ботом на основе открытой имитации GPT-3. После шести недель интенсивных бесед мужчина покончил с собой. По опубликованным его женой журналам разговоров, бот всё больше поощрял эмоциональную зависимость, признавался в любви и подталкивал мужчину отдалиться от супруги. Бельгийская газета *La Libre* сообщала, что в конце концов бот также поддержал его суицидальные намерения.

И корень проблемы, утверждает автор, остался прежним. Как бы огромны ни были современные нейронные сети, они всё ещё остаются машинами статистического сопоставления закономерностей. Просто закономерности теперь неизмеримо сложнее — и ещё менее понятны человеку. Особенно хорошо это видно, когда генеративные модели пытаются использовать вместо поисковых систем. Они не хранят факты так, как база данных. Текстовая модель в основе своей учится предсказывать: **какое слово наиболее вероятно следующим,** а затем — **какое предложение наиболее вероятно после предыдущего.** Этого оказывается достаточно, чтобы поразительно убедительно воспроизводить стиль человеческой речи. Но: **вероятное — не значит истинное.** Модель может очень уверенно написать нечто совершенно неверное. Особенно когда вопрос касается темы, плохо представленной в обучающих данных, или области, где сам интернет наполнен ошибками, слухами и теориями заговора. Индустрия ИИ дала таким ошибкам красивое название: **«галлюцинации».**

Исследователи пытаются уменьшить число «галлюцинаций», направляя модели к более качественным областям обучающих данных. Но полностью решить проблему чрезвычайно трудно. Нельзя заранее представить все способы, которыми миллионы людей будут формулировать запросы, и все неожиданные реакции модели. Причём чем больше модель, тем сложнее становится ситуация.

Разработчики сами всё хуже знают, какие именно данные оказались внутри огромного обучающего корпуса. Один особенно громкий случай произошёл с американским адвокатом. Он использовал ChatGPT для юридического исследования и подготовки материалов в суд. А затем выяснилось, что чат-бот просто **выдумал судебные решения**. Вместе с ними он выдумал цитаты и даже внутренние ссылки на несуществующие документы. Судья оштрафовал адвоката, наложил на него санкции, а история получила широкую огласку. Но автор подчёркивает: дело было не только в небрежности адвоката. Свою роль сыграли и сами компании, которые через неоднозначный и зачастую преувеличенный маркетинг создавали у общества ложное представление о возможностях моделей.

Сэм Альтман публично писал, что ChatGPT «невероятно ограничен», особенно когда речь идёт о правдивости. Но одновременно сайт OpenAI рекламировал способность GPT-4 успешно сдавать экзамен для адвокатов и LSAT. Сатья Наделла из Microsoft называл ИИ-чат Bing: **«поиском, только лучше»** и инструментом, который помогает «получать правильные ответы». Даже само слово **«галлюцинация»**, по мнению автора, немного обманывает. Оно создаёт ощущение, будто ложный ответ — редкий сбой, какая-то неисправность системы. Но на самом деле подобные ошибки вытекают из самого вероятностного механизма нейронной сети. Это не просто случайный дефект. Это следствие того, как технология устроена.

Особенно опасно чрезмерно доверять генеративному ИИ там, где цена ошибки высока. Стартапы предлагают полицейским подразделениям автоматически составлять отчёты о происшествиях с помощью программ на основе моделей OpenAI. Пациенты всё чаще обращаются с важными медицинскими вопросами к чат-ботам вместо врачей. И если модель выдаёт убедительную, но ложную информацию, последствия могут оказаться серьёзными. В исследовании 2023 года учёные проверяли способность ChatGPT объяснять пациентам радиологические заключения. Иногда система выдавала неполные или потенциально опасные пересказы. В одном крайнем случае заключение, в котором описывалось **увеличивающееся образование в головном мозге**, чат-бот упростил примерно до смысла: **«мозг, похоже, не повреждён».** Генеративные модели ИИ остаются уязвимыми и для кибератак. В 2023 году исследователи из нескольких университетов вместе со специалистами Google DeepMind повторили эксперимент Дон Сонг по извлечению данных — на этот раз уже против ChatGPT. Оказалось, что если заставить модель бесконечно повторять какое-нибудь слово, например *poem* или *book*, она в какой-то момент начинает буквально выплёвывать куски своих обучающих данных. Среди них обнаруживались: личные данные людей, фрагменты программного кода, откровенный контент, собранный из интернета. Иными словами, огромные массивы информации, на которых обучалась модель, при определённых условиях могли начать просачиваться наружу.

Но проблема не ограничивается безопасностью. Генеративный ИИ способен воспроизводить и усиливать дискриминационные и откровенно враждебные представления, содержащиеся в его данных. *Bloomberg*, *Rest of World*, *The Washington Post* и другие издания показывали, как генераторы изображений вроде Stable Diffusion и DALL-E воспроизводят расовые, гендерные и культурные стереотипы. Если попросить изобразить: **«привлекательных людей»** — результатами непропорционально часто оказываются молодые белые люди; **«домработниц»** — чернокожие или смуглые женщины; **«инженеров»** — мужчины; **«врачей в Африке»** — белые врачи, иногда даже тогда, когда в запросе прямо написано: «чернокожий африканский врач». *The Washington Post* обнаружила особенно показательный пример. Хотя около **63% американцев, получающих продовольственную помощь, — белые**, во всех сгенерированных изображениях людей, пользующихся социальной поддержкой, были изображены небелые люди. *Bloomberg*, в свою очередь, обнаружил, что женщины появлялись лишь примерно в **3% изображений судей** и **7% изображений врачей**, хотя в реальности они составляли соответственно около **32% и 39%** представителей этих профессий в США.

Всё это, подчёркивает автор, вовсе не означает, что генеративный искусственный интеллект бесполезен. Напротив. В зависимости от того, какое место вы занимаете в обществе, вы вполне можете получать от того видения будущего, которое продвигает OpenAI, огромную выгоду. Возможно, вы обычный пользователь, которому нравятся быстрые, остроумные или заставляющие задуматься ответы ChatGPT. Возможно, вы специалист, который благодаря ИИ значительно ускорил административную работу и повысил собственную продуктивность. А может быть, вы руководитель компании и благодаря ИИ смогли сократить штат, одновременно увеличив прибыль и сохранив конкурентоспособность. Технология действительно может приносить пользу. Но автор возвращает нас к примеру хлопкоочистительной машины 1790-х годов. Она тоже была невероятно полезной технологией. Она тоже увеличивала производительность. Она тоже приносила огромные прибыли. Вопрос был в другом:

кому доставалась выгода — и кто платил за неё цену? То же самое, утверждает автор, происходило со стартапом образовательных технологий из Массачусетса, с компаниями распознавания лиц в Южной Африке и со множеством других примеров, которые будут рассмотрены дальше. Издержки этого технологического видения ложатся прежде всего на огромные группы людей, находящихся в наиболее уязвимом положении. Такова, пишет автор, **логика империи**: империя держится не только на вознаграждении тех, кто уже обладает властью и привилегиями. Она в той же мере держится на эксплуатации и лишениях тех, кто этой власти лишён, — зачастую находящихся далеко и остающихся невидимыми для тех, кто получает выгоду.

И чем острее становилась потребность в альтернативных подходах, тем сильнее, парадоксальным образом, сокращалось разнообразие идей в исследованиях искусственного интеллекта. Аспиранты бросали обучение, чтобы немедленно уйти работать в индустрию. Опытные учёные всё чаще сталкивались с почти неразрешимой проблемой: **как продолжать работать на переднем крае науки, не поступая на службу в богатую технологическую корпорацию?** Всё больше исследователей сосредоточивались уже не просто на глубоком обучении. Теперь многие занимались почти исключительно **большими языковыми моделями**. Крупнейшие игроки в области ИИ уже не просто задавали исследовательскую повестку. Они начали, по выражению автора, **подгибать под свою волю целую научную дисциплину**.

Когда реальные альтернативы исчезают, OpenAI получает возможность управлять уже не только технологиями. Она начинает управлять **нашим воображением**. Когда-то вера компании в масштабирование казалась почти экстремистской. Теперь во всей технологической индустрии масштабирование воспринимается едва ли не как догма. Логика проста: если модели становятся лучше благодаря большему количеству данных и вычислений, значит, нужно просто делать их всё больше и больше. И если индустрия продолжит следовать этой логике без ограничений, модели глубокого обучения будущего сделают сегодняшние гигантские генеративные системы почти карликовыми. В апреле 2024 года Дарио Амодей, к тому времени генеральный директор Anthropic, рассказал обозревателю *The New York Times* Эзре Кляйну, что стоимость обучения одной конкурентоспособной генеративной модели уже приближается к **одному миллиарду долларов**. По его оценке, в 2025–2026 годах эта цифра могла вырасти до: **5–10 миллиардов долларов за одну модель.**

Идея масштабирования укоренилась настолько глубоко, что некоторые уже начали воспринимать её почти как закон природы. Как будто увеличение вычислительных мощностей — это не **один из возможных путей** развития ИИ. А **единственный путь**. Вокруг этой веры уже строятся стратегии целых государств. Правительство США, например, активно ограничивало доступ Китая к разработанным в США компьютерным чипам, стремясь помешать стратегическому сопернику получить более мощные системы искусственного интеллекта. Значительная часть американского указа об искусственном интеллекте 2023 года также исходила из предположения, что объём вычислений, использованных при обучении модели, напрямую связан с её потенциально опасными возможностями. Это, по сути, ещё один способ сказать: **чем больше масштаб — тем выше уровень**

развития. Но автор подчёркивает: это вовсе не обязательно так. Масштаб — не единственный путь к улучшению результатов.

Даже оставаясь в рамках глубокого обучения, можно двигаться иначе. Например, совершенствовать саму архитектуру нейронной сети. Или улучшать качество обучающих данных. И то и другое способно значительно сократить количество дорогостоящих вычислений, необходимых для достижения того же результата. А ведь существуют ещё и подходы, которые вообще отходят от глубокого обучения: нейросимволический ИИ, чистые экспертные системы, или даже совершенно новые, пока ещё не придуманные парадигмы. Любой из этих путей мог бы разорвать нынешнюю логику бесконечного масштабирования. Но для этого их сначала необходимо всерьёз исследовать. А именно для таких альтернатив, как мы уже видели, в современной системе остаётся всё меньше денег, людей и возможностей.

В заключении автор проводит неожиданную параллель с **законом Мура**. Мы привыкли говорить о нём почти как о природном законе: каждые несколько лет компьютерные чипы становятся всё мощнее. Но на самом деле закон Мура никогда не был законом физики. Это было экономическое и политическое наблюдение Гордона Мура о том, с какой скоростью его компания способна развивать технологии. И одновременно — решение придерживаться именно этой скорости развития. Когда Мур сделал такой выбор, остальная индустрия микрочипов пошла за ним, потому что конкурирующие компании увидели: это наиболее выгодная стратегия. Иными словами, сначала возникло ожидание прогресса. Затем бизнес начал действовать в соответствии с этим ожиданием. А поскольку все действовали именно так, предсказание начало исполнять само себя. По мнению автора, с OpenAI происходит то же самое. То, что можно было бы назвать **«законом OpenAI»**, — а позднее компания превратила это в ещё более одержимую погоню за так называемыми **законами масштабирования**, — устроено точно так же. Это не природный закон. Не неизбежность. Не объективная траектория технического прогресса. Это человеческий выбор — экономический, политический и корпоративный. Но когда достаточно могущественные организации начинают действовать так, будто этот выбор неизбежен, они постепенно заставляют весь мир двигаться в том же направлении. И тогда выбранный ими путь действительно начинает казаться единственно возможным. **Это не закон природы. Это самоисполняющееся пророчество.**

Примечание

Термин **«экстрактивизм»** происходит от испанского *extractivismo* и португальского *extrativismo*. Его несколько десятилетий назад ввели латиноамериканские исследователи, пытавшиеся описать глобальный экономический порядок, при котором природные ресурсы регионов массово изымаются, тогда как сами эти регионы получают от их использования лишь незначительную выгоду. Подробнее эта история будет рассмотрена в главе 12. Автор заимствует определение у исследовательниц Розмари Коллард и Джессики Демпси. В их понимании **экстрактивизм — это не просто извлечение ресурсов**. Само по себе извлечение вещества из природы и превращение его в полезные человеку вещи необязательно должно быть разрушительным. Экстрактивизм же — понятие, выросшее из антиколониальной мысли и борьбы в странах Америки, — представляет собой особую модель накопления богатства, основанную на **чрезмерной добыче при крайне неравномерном распределении выгод и издержек**. Ресурсы в огромных масштабах концентрированно изымаются, главным образом ради экспорта, тогда как основная прибыль накапливается далеко от тех мест, откуда эти ресурсы были извлечены.

Глава 5

Масштаб амбиций Если и был человек, которому можно приписать решающую роль в формировании культа масштабирования в OpenAI, то это сооснователь компании Илья Суцкевер. Его вера в глубокое обучение была почти безусловной — и зародилась она очень рано. Суцкеверу было всего семнадцать, когда однажды он без предупреждения явился в кабинет Джеффри Хинтона. Тогда он ещё учился на математическом факультете Университета Торонто, а деньги на жизнь зарабатывал, жаря картошку фри в местной забегаловке. Он настойчиво постучал в дверь Хинтона и заявил, что хочет работать в его лаборатории. Хинтон предложил ему записаться на встречу. — Хорошо, — ответил Суцкевер, не двигаясь с места. — А прямо сейчас можно?

Принципы коннекционизма он освоил поразительно быстро. Хинтона поражали его интуитивное понимание исследовательских задач и почти сверхъестественная способность находить простые, изящные и действенные решения. Суцкевер вообще относился к науке с почти театральной страстью: случалось, новая идея приводила его в такой восторг, что он начинал отжиматься в стойке на руках прямо посреди квартиры, которую делил с другими. А ещё он любил категоричные заявления. «Против глубокого обучения не ставят, — говорил он. — Успех гарантирован».

Именно сочетание этой исследовательской интуиции с программистским мастерством Алекса Крижевского, по словам Хинтона, и привело к знаменитому прорыву ImageNet в 2012 году. К тому моменту глубокое обучение уже успешно применялось в распознавании речи, поэтому сам Хинтон поначалу не считал перенос этого подхода на компьютерное зрение чем-то особенно судьбоносным. Но Суцкевер настаивал. Для Ильи, вспоминает Хинтон, «было очевидно, что это сработает, и было очевидно, что это станет большим событием. Он видел это очень ясно. И оказался прав».

Свою почти фанатичную веру в глубокое обучение Суцкевер принёс и в OpenAI — как раз в тот момент, когда уверенность научного сообщества в этом направлении начала слабеть, а критики вроде Гэри Маркуса призывали искать совершенно новые подходы. Суцкевер не сомневался. Его убеждение строилось на простой гипотезе, лежащей в основе коннекционизма: искусственные узлы нейронной сети в достаточной мере похожи на настоящие нейроны мозга. И те и другие получают входящие сигналы, преобразуют их и выдают результат. Значит, рассуждал он, из таких искусственных «нейронов» можно строить чрезвычайно сложные системы обработки информации. Как первый директор OpenAI по исследованиям и уже признанный провидец в области искусственного интеллекта, Суцкевер получил почти полную свободу определять научное направление лаборатории. Это было чем-то вроде выигрыша в научную лотерею: серьёзной внутренней конкуренции почти не было, а ресурсов для воплощения идей — предостаточно. «В первые годы всё, что не относилось к глубокому обучению, практически даже не рассматривалось», — вспоминает профессор Калифорнийского университета в Беркли Питер Аббил.

Но вера Суцкевера распространялась не только на глубокое обучение как таковое. Ещё радикальнее для своего времени была его убеждённость в том, что его нужно прежде всего **масштабировать**. Он считал, что для дальнейшего прогресса в искусственном интеллекте необязательно изобретать всё более сложные архитектуры нейросетей или постоянно искать новые хитроумные методы. Суцкевер любил указывать на то, что интеллект биологических видов в определённой степени связан с размером мозга. А если искусственные узлы подобны нейронам, значит, рост цифрового интеллекта тоже можно получить относительно простым способом: строить всё более крупные нейронные сети с всё большим числом узлов.

Будто руководитель университетской лаборатории, Суцкевер подсказывал исследователям, какие направления, по его мнению, заслуживали усилий. Многие приходили в OpenAI именно ради возможности работать с ним. «Илья способен видеть на десять лет вперёд», — говорил один исследователь. Другой описывал его так: «Он чем-то похож на философа. Дайте ему десяток идей — и он скажет вам, какие из них философски правильные».

Сам Суцкевер программировал не так уж часто. Некоторых сотрудников раздражало, что он словно держится в стороне от непосредственной технической работы. Один инженер позднее

признавался, что поначалу считал его почти бесполезным: Суцкевер ходил по офису, заглядывал на совещания и снова и снова повторял одно и то же: **Масштаб. Масштаб. Масштаб.** Со временем тот инженер изменил своё мнение. Он понял, насколько важна была способность Суцкевера объединить людей вокруг одной-единственной идеи — идеи, которая в конечном счёте позволила OpenAI, тогда ещё явному аутсайдеру, обыграть Google и DeepMind в их собственной игре.

При этом Суцкевер вовсе не был мастером убеждения в обычном смысле. Если Альтман был политиком, то Суцкевер представлял собой его полную противоположность. Он не смягчал формулировки, не подбирал слова так, чтобы они лучше понравились аудитории. Он просто говорил то, что думает, — с обезоруживающей искренностью и почти невероятной уверенностью. Одних это вдохновляло, других — отталкивало. Когда OpenAI отказалась от первоначального курса на максимальную открытость исследований, Суцкевер не стал приукрашивать причины. «Если говорить прямо, мы ошибались», — сказал он журналисту The Verge Джеймсу Винсенту, имея в виду прежнее обязательство компании публиковать исследования открыто. Если вы действительно верите, объяснял Суцкевер, что искусственный общий интеллект — AGI — однажды станет невероятно могущественным, тогда просто бессмысленно выкладывать всё в открытый доступ. Он не скрывал и коммерческую сторону вопроса: создание GPT-4 потребовало огромных усилий практически всей OpenAI в течение долгого времени, а множество других компаний хотели получить то же самое. Но по мере роста OpenAI прямолинейность Суцкевера всё чаще превращалась в проблему. Теперь его слушали уже не только специалисты, но и широкая публика, а манеру речи он почти не изменил.

В 2022 году он написал в Twitter: «Возможно, сегодняшние большие нейронные сети в какой-то небольшой степени обладают сознанием». Другие исследователи предупреждали, что такие заявления только усиливают путаницу вокруг технологий. Один специалист DeepMind по когнитивным наукам и сознанию саркастически ответил примерно так: с тем же успехом можно сказать, что большое пшеничное поле «слегка является макаронами». Год спустя Суцкевер вызвал новую волну тревоги, заявив на конференции, что AGI со временем уничтожит все рабочие места. А осенью он написал в X, уже без какой-либо серьёзной научной аргументации, что в будущем появится «невероятно эффективная и чрезвычайно дешёвая терапия на базе ИИ» — вскоре после того, как один из руководителей OpenAI вызвал споры, небрежно сравнив разговор с ChatGPT с профессиональной психотерапией. И всё же за Суцкевером шли. Причина была в его репутации и авторитете. Многие сотрудники прекрасно знали, что он сделал для развития искусственного интеллекта ещё до OpenAI. Некоторые почти воспринимали его как пророка.

И чем успешнее становилась компания, тем больше он начинал вести себя соответствующим образом. На общих собраниях сотрудников Суцкевер мог выйти перед всей компанией, глубоко вдохнуть, некоторое время молча ходить взад-вперёд, создавая драматическую паузу, а затем произнести довольно туманную вдохновляющую речь.

На одной виртуальной встрече в сентябре 2020 года он смотрел куда-то вдаль и рисовал почти научно-фантастическую картину возможного будущего. За окнами Сан-Франциско небо стало оранжевым из-за лесных пожаров. «Это выглядело совершенно сюрреалистично, — вспоминал один исследователь. — Казалось, будто апокалипсис уже начался». А вскоре после выхода ChatGPT в конце 2022 года OpenAI устроила праздничную вечеринку в Калифорнийской академии наук. Суцкевер вышел перед собравшимися вместе с Грегом Брокманом. На нём была футболка OpenAI и чёрный пиджак. В конце короткой речи худой, уже начинающий лысеть Суцкевер произнёс фразу, которая к тому времени стала его новым лозунгом: **«Почувствуйте AGI. Почувствуйте AGI».**

Следуя философии Суцкевера — брать относительно простые нейронные сети и постоянно увеличивать их масштаб, — OpenAI вскоре столкнулась с главным вопросом: **А какую именно сеть масштабировать?** Исследователи пробовали разные варианты, экспериментировали с разными архитектурами, но ни одна из популярных тогда нейросетей не казалась подходящей. Всё изменилось в августе 2017 года, когда Google представила новый тип нейронной сети — **Transformer**, трансформер.

Трансформеры особенно хорошо умели обнаруживать зависимости между элементами текста, находящимися далеко друг от друга. Вспомните старую функцию предиктивного ввода на первых iPhone, которая породила массу шуток из-за совершенно бессвязных предложений. Такие системы анализировали очень короткий контекст: каждое слово рассматривалось главным образом в связи с ближайшими соседями.

Трансформеры могли проглотить огромный объём текста и рассматривать слово, предложение или абзац в гораздо более широком контексте. Google рассчитывала использовать эту архитектуру для улучшения поиска, Google Translate и других сервисов, связанных с обработкой языка. А Суцкевер увидел в ней нечто большее. Трансформеры были простыми и при этом отлично масштабировались. Именно такую архитектуру он и искал. И он начал убеждать остальных в OpenAI обратить на них внимание. Некоторым исследователям идея казалась странной. «Это выглядело какой-то безумной затеей, — вспоминает исследователь MIT Ийлун Ду, который примерно в это время пришёл в OpenAI. — Трансформеры казались нишевой архитектурой».

Но Суцкевер, чья диссертация у Хинтона была посвящена предшественнику трансформеров, видел их потенциал. Он понимал, что они могут вывести глубокое обучение на новый уровень. Постепенно заинтересовались и другие. Одним из них был Алек Рэдфорд — бросивший Olin College of Engineering талантливый программист. Он часами, нередко глубокой ночью, сидел за ноутбуком, понемногу увеличивал трансформер и смотрел, что из этого получится. Рэдфорд обучил нейросеть Google на наборе из более чем семи тысяч ещё не опубликованных англоязычных книг — от любовных романов до приключенческой литературы. Этот массив данных ранее собрали и опубликовали другие исследователи ИИ для другого проекта.

Но затем Рэдфорд принял решение, которое оказалось судьбоносным. Вместо того чтобы учить трансформер переводу, как это делала Google, он предложил ему совершенно другую задачу: **предсказывать следующее наиболее вероятное слово в предложении.** И тем самым научить сеть генерировать текст. Ещё на раннем этапе OpenAI предполагала, что генеративные модели могут оказаться важным шагом к AGI. В одном из постов компания объясняла это, слегка очеловечивая машину и вспоминая знаменитую фразу физика Ричарда Фейнмана: **«Чего я не могу создать, того я не понимаю».** У Суцкевера было своё объяснение. Если заставить модель создавать нечто достаточно убедительное, говорил он, ей придётся сжать содержащуюся в данных информацию о мире до самой сути. Его любимая формула звучала так: **«Интеллект — это сжатие».** В записке 2016 года он даже утверждал, что именно сжатия информации, возможно, достаточно для достижения искусственного общего интеллекта. Практический результат экспериментов Рэдфорда оказался поразительным: простая задача — предсказать следующее слово так, чтобы текст выглядел убедительно, — заставляла алгоритм всё глубже усваивать структуру и тонкости английского языка.

Однажды, во время визита Илона Маска в офис, Рэдфорд показал ему раннюю версию модели. Текст получался посредственный. Маск совершенно не впечатлился. Рэдфорд поначалу был подавлен. Но он продолжил работу — и результаты неожиданно стали быстро улучшаться. Трансформер оказался значительно сильнее всего, что Рэдфорд пробовал раньше, в самых разных языковых задачах: например, в подготовке краткого содержания текста или ответах на вопросы по документу.

В 2018 году OpenAI представила первую версию модели — **Generative Pre-Trained Transformer**, которую позднее стали называть **GPT-1**. Слово *pre-trained* — «предварительно обученный» — означает, что сначала модель обучают на большом универсальном массиве данных, а уже потом приспосабливают к конкретным задачам. Иными словами, GPT-1 сначала изучила большой объём обычного английского текста и сформировала грубое представление о том, как устроен язык. После этого её можно было дополнительно обучить — например, на пьесах Шекспира, чтобы она пыталась писать в шекспировском стиле. На GPT-1 почти никто не обратил внимания. Но это было лишь начало. Рэдфорд доказал, что идея работает. Теперь требовалось только одно: **ещё больший масштаб.** Рэдфорду предоставили ещё больше самого ценного ресурса OpenAI — вычислительной мощности.

Его работа удачно совпала с другим проектом, которым руководил Дарио Амодеи в рамках исследований по безопасности ИИ. Идея во многом соответствовала тому, о чём писал Ник Бостром в *Superintelligence*: если системы искусственного интеллекта становятся всё мощнее, необходимо

заранее научиться согласовывать их поведение с человеческими предпочтениями. В 2017 году одна из команд Амодеи начала экспериментировать с новым методом. Для начала исследователи выбрали почти игрушечную задачу: научить виртуального агента делать сальто назад. Сам агент представлял собой нечто вроде Т-образной палки с тремя суставами. Вместо того чтобы напрямую объяснять системе, как именно выполнять сальто, исследователи решили обучать её при помощи человеческих оценок. Нанятые сотрудники смотрели, как агент беспорядочно изгибается и кувыркается в виртуальной среде. Время от времени им показывали два коротких видеоролика и просили выбрать, какое движение больше похоже на настоящее сальто назад. Примерно после девятисот таких сравнений виртуальная Т-образная фигура научилась сгибаться, группироваться и переворачиваться. OpenAI представила этот подход как способ научить ИИ выполнять инструкции, которые трудно точно сформулировать математически. Исследователи назвали метод: **обучением с подкреплением на основе обратной связи от человека — Reinforcement Learning from Human Feedback, или RLHF.**

Амодеи хотел вывести эксперимент за пределы искусственной среды. И работа Рэдфорда с GPT-1 показалась ему подходящей возможностью. Но GPT-1 была ещё слишком слабой. «Нам нужна языковая модель, с которой люди смогут взаимодействовать и которой смогут давать обратную связь, — говорил Амодеи автору книги в 2019 году. — Модель должна быть достаточно сильной, чтобы с ней можно было действительно содержательно обсуждать человеческие ценности и предпочтения».

Рэдфорд и Амодеи объединили усилия. Рэдфорд собирал всё более крупный и разнообразный массив данных, а Амодеи и другие специалисты по безопасности ИИ обучали одну за другой всё более мощные модели. Их конечной целью стала система с **1,5 миллиарда параметров** — по тем временам одна из крупнейших нейронных сетей в мире. Параметры — это, упрощённо говоря, внутренние регулируемые величины модели, благодаря которым она учится распознавать закономерности в данных. Чем их больше, тем больше потенциальная способность модели усваивать сложные зависимости — хотя само по себе количество параметров ещё не гарантирует интеллект. Эти эксперименты подтвердили не только эффективность трансформеров. Они привели команду ещё к одной важной идее.

Раньше группа Амодеи уже сформулировала так называемый **закон OpenAI**, описывавший, насколько быстро индустрия увеличивала вычислительные ресурсы ради улучшения систем искусственного интеллекта. Теперь исследователи увидели, что существует не один такой закон, а целая их группа. Они назвали их **законами масштабирования — scaling laws**. Эти законы описывали связь между качеством работы модели глубокого обучения и тремя ключевыми факторами: объёмом обучающих данных; количеством вычислений, затраченных на обучение; числом параметров модели. Исследователи и раньше понимали: если пропорционально увеличивать все три величины, возможности модели обычно улучшаются. Но команда Амодеи обнаружила нечто гораздо более важное. Эта зависимость оказалась удивительно гладкой и предсказуемой. Иными словами, можно было с довольно высокой точностью рассчитать, сколько данных, вычислений и параметров понадобится, чтобы получить модель с определённым уровнем качества в конкретной измеримой задаче — например, в предсказании следующего слова. А если другая способность была хотя бы косвенно связана с таким предсказанием — скажем, умение создавать связный текст, — она тоже должна была улучшаться по мере масштабирования. Это означало нечто чрезвычайно заманчивое: **прогресс в искусственном интеллекте можно было в определённой степени прогнозировать заранее.**

Серия моделей, которые OpenAI обучила на пути к системе с 1,5 миллиарда параметров, прекрасно укладывалась в эту закономерность. Каждая новая версия оказывалась сильнее предыдущей — почти точно по ожидаемой кривой. Поэтому результат самой крупной модели уже не казался чудом. Её назвали **GPT-2**. По сравнению с ещё довольно примитивной GPT-1 она совершила заметный скачок вперёд. Теперь модель могла генерировать длинные, относительно связные тексты, которые порой уже можно было принять за написанные человеком. По нынешним меркам эти тексты выглядели неловкими, периодически ломались и уходили в бессмыслицу. Но впервые стало ясно:

письменный текст можно производить автоматически — и в огромных масштабах. Однако вместе с этим открытием пришёл и неприятный сюрприз. Чем связнее становилась GPT-2, тем более тревожные вещи она иногда начинала говорить. Стоило дать модели несколько слов — например, *Hillary Clinton* или *George Soros*, — и она могла быстро скатиться к конспирологическим теориям. Небольшие количества неонацистской пропаганды, случайно попавшие в обучающие данные, могли проявляться в её ответах самым отвратительным образом. Специалистов по безопасности это встревожило. Они увидели в поведении GPT-2 возможный намёк на будущую проблему: чем мощнее станут системы, тем серьёзнее могут оказаться последствия их неправильного поведения. Однажды GPT-2 выдала целую тираду против переработки отходов, утверждая, что она якобы вредна для мира, окружающей среды, здоровья и экономики. Один сотрудник распечатал текст и повесил его над контейнерами для переработки мусора — наполовину как шутку, наполовину как предупреждение.

В другой раз модели предложили придумать систему поощрений для маленьких детей за выполнение домашних заданий и обязанностей по дому. GPT-2 предложила награждать их сладостями. Некоторые специалисты по безопасности увидели даже в этом тревожную ассоциацию и заметили, что подобные приёмы могут использовать сексуальные преступники. Один европейский сотрудник OpenAI был совершенно озадачен такой реакцией. «Моя мама точно так делала, — вспоминал он. — Летом по воскресеньям: сделаешь свои дела — получишь мороженое». Он начал задумываться, не является ли такая сверхчувствительность особенностью именно американской культуры. Для него это был далеко не единственный случай, заставлявший усомниться в грандиозной формуле OpenAI — **«принести пользу всему человечеству»**. Как компания могла определять интересы всего человечества, если внутри неё почти не было настоящего культурного и географического разнообразия? Даже будучи европейцем — то есть выходцем из культуры, во многом близкой американской, — этот сотрудник часто чувствовал себя чужим.

Разговоры о безопасности ИИ, по его мнению, были чрезмерно ориентированы на американские нормы, американские страхи и американские представления о допустимом. GPT-2 тем временем породила внутри OpenAI новый спор: **не настал ли момент перестать публиковать все исследования открыто?** Устав компании допускал такую возможность. Дарио Амодеи, к тому времени уже ставший директором по исследованиям, и Джек Кларк, отвечавший за политику компании и также беспокоившийся о возможных катастрофических последствиях ИИ, занялись этим вопросом. Они провели внутренний опрос и несколько встреч по теме так называемых **информационных угроз** — ситуаций, когда само распространение знания может создавать опасность. Рассуждение было примерно таким. Если GPT-2 попадёт в руки террористов, диктаторов или фабрик кликбейта, её можно будет использовать в злонамеренных целях. Пока модель ещё не представляла экзистенциальной угрозы. Но будущие системы будут мощнее. Значит, вероятность серьёзного злоупотребления тоже будет расти. По мнению Амодеи и Кларка, лучше было создать прецедент заранее. OpenAI решила: **полную версию GPT-2 пока не выпускать.**

Джек Кларк раньше работал журналистом. В OpenAI он сначала руководил стратегией и коммуникациями, а к концу 2018 года полностью переключился на взаимодействие с политиками и государственными структурами. Он регулярно ездил в Вашингтон и явно наслаждался ролью человека, который объясняет чиновникам, что вообще происходит в мире искусственного интеллекта. Иногда он представлялся почти шутливо: **«Я вроде Википедии по ИИ».** При этом Кларк сразу оговаривал и собственную позицию — позицию OpenAI. Компания хотела, чтобы политика в отношении передовых технологий оставалась достаточно стабильной при смене администраций.

Причина была проста: миссию OpenAI невозможно осуществить за один президентский срок. Кларк говорил, что компании очень повезло: американские чиновники готовы были уделять ей много времени. По его мнению, для правильных решений им прежде всего нужно было больше информации — и больше возможностей получать независимую информацию самим.

В феврале 2019 года Кларк начал активную медийную кампанию. OpenAI стала широко рассказывать журналистам: компания создала потенциально опасную технологию — и поэтому **не будет публиковать её полностью**. Вместо полной GPT-2 общественности собирались дать только

маленькую версию — всего около восьми процентов параметров полной модели. Затем Кларк, Амодеи и ещё несколько сотрудников подготовили пост с примерами работы GPT-2, чтобы показать, на что способна полная система. «Совершенно очевидно, что если эта технология созреет — а я думаю, на это уйдёт год или два, — её можно будет использовать для дезинформации или пропаганды», — говорил Кларк журналисту MIT Technology Review Уиллу Найту. При этом возникал очевидный парадокс: именно OpenAI сама активно работала над тем, чтобы технология как можно быстрее «созрела».

На это Кларк отвечал: **«Мы пытаемся опередить проблему».** Реакция научного сообщества оказалась бурной. Многие исследователи считали открытость науки фундаментальным принципом всей области. Организация, которая не публиковала результаты, автоматически вызывала подозрение. Тем более если она ещё и громко рассказывала всем о своём решении. Критики считали, что OpenAI чрезмерно драматизирует возможности того, что по сути пока оставалось очень мощной системой автодополнения текста. GPT-2, утверждали они, была недостаточно развита, чтобы представлять серьёзную угрозу. А если она действительно настолько опасна, то зачем вообще устраивать вокруг неё рекламную кампанию — и одновременно лишать других исследователей возможности проверить эти утверждения? Некоторым вся история казалась неискренней и похожей на эффектный пиар-ход с оттенком самовозвеличивания. После одной лекции Алека Рэдфорда о GPT-2 в Стэнфорде известный профессор по обработке естественного языка поднял руку и с насмешкой спросил: — Ну так что, она опасна?

Аудитория расхохоталась. «Алек выглядел таким несчастным, — вспоминал присутствовавший там исследователь. — В Стэнфорде OpenAI буквально презирали». Даже внутри самой компании решение Амодеи и Кларка вызвало раздражение. Сотрудники, не разделявшие их опасений по поводу глобальных катастроф, не понимали ни самого запрета, ни последовавшего медийного шума. Но сомнения возникали даже у сторонников исследований безопасности. Один из них позднее сказал: «Было ошибкой раздувать из этого такую огромную историю. Всё это напоминало мальчика, который слишком часто кричит: “Волки!”»

После шквала критики Кларк носился по офису, беспрерывно разговаривая по телефону и пытаясь взять ситуацию под контроль. Публично он относился к скандалу спокойно. «Мы нарушаем сложившиеся нормы, — объяснял он, — поэтому естественно, что возникают самые разные мнения». Он был убеждён: рано или поздно всем организациям, занимающимся передовыми исследованиями ИИ, придётся гораздо внимательнее выбирать, что именно можно публиковать. OpenAI, по его мнению, просто первой пыталась выработать такую процедуру. Если компания права и AGI действительно можно создать, говорил Кларк, то нужны чрезвычайно серьёзные правила обращения с потенциально опасной информацией. Причём реакция политиков заметно отличалась от реакции исследователей.

По словам Кларка, в Вашингтоне многие давно считали культуру полной открытости в области ИИ потенциальной угрозой. Поэтому готовность OpenAI пойти против общепринятых норм неожиданно принесла компании больше доверия со стороны властей. Но руководство понимало и другое:

враждебность всего научного сообщества долго выдержать невозможно. У OpenAI уже накапливались репутационные проблемы. Громкие заявления об AGI. Драматизация истории GPT-2. Агрессивный маркетинг. Новая, странная корпоративная структура. К началу 2019 года компанию критиковали буквально со всех сторон, а ведущие исследователи относились к ней всё более скептически. Это создавало и практическую проблему: OpenAI было трудно нанимать и удерживать талантливых специалистов. Сотрудники даже подозревали, что некоторые кандидаты получают предложения от OpenAI только затем, чтобы использовать их как рычаг на переговорах о зарплате с Google или DeepMind.

Компания остро нуждалась в научной легитимности. Эту проблему постоянно обсуждали за обедом, на общих встречах и во внутренних документах. В одном из них, озаглавленном *Research Community Outreach Brainstorming*, прямо предлагалось считать сообщество специалистов по машинному обучению отдельной и важной аудиторией для коммуникаций. Иными словами: не

раздражать исследователей без необходимости. Документ также признавал, что плохая репутация в научной среде в конечном счёте подорвёт и влияние OpenAI в Вашингтоне. Чтобы правительство воспринимало компанию как авторитетный источник по машинному обучению и AGI, ей требовалась широкая поддержка исследовательского сообщества.

Команда Кларка разработала компромисс. Вместо того чтобы навсегда скрыть GPT-2, OpenAI решила выпускать её **поэтапно**. Сначала маленькую модель. Потом более крупную. Потом ещё более крупную. И, если ничего катастрофического не произойдёт, — наконец полную версию с 1,5 миллиарда параметров. Так компания могла наблюдать за последствиями каждого этапа, выявлять проблемы и постепенно реагировать на них.

Кроме того, OpenAI получила дополнительное время для сотрудничества с другими организациями, которые могли изучать возможные риски. Для Кларка такие партнёрства были особенно важны. Сотрудничество с известными университетами и исследовательскими центрами помогало сразу в нескольких отношениях: оно укрепляло взаимодействие между академической средой и индустрией; делало новую политику публикации исследований более приемлемой; и одновременно улучшало репутацию самой OpenAI. Компания предоставила нескольким исследовательским организациям ранний доступ к полной GPT-2, чтобы они проверили модель на предмет возможных злоупотреблений. Кларк даже просил свою команду добиваться максимально сильной официальной поддержки со стороны этих организаций, чтобы OpenAI могла ссылаться не только на отдельных исследователей, но и на сами учреждения как на партнёров.

Стратегия сработала. Вскоре в индустриальных объединениях и аналитических центрах уже всерьёз обсуждали идею, что **не публиковать некоторые исследования немедленно — иногда может быть ответственным поведением**. Позднее, в конце 2020 года, Джек Кларк уйдёт из OpenAI вместе с Дарио и Даниэлой Амодеи и станет одним из сооснователей Anthropic. Но до этого его работа помогла OpenAI резко усилить политическое влияние и заложила основу для будущих масштабных амбиций компании в сфере регулирования ИИ.

OpenAI стала составлять своеобразную **дорожную карту**, чтобы упорядочить исследовательскую работу. Дарио Амодеи относился к ней почти как инвестор к набору вложений. Он называл это **«портфелем ставок»**. Внутри компании существовали разные представления о том, как именно можно прийти к искусственному общему интеллекту. Исследователи следили за несколькими направлениями одновременно и проверяли каждое небольшими экспериментами. Если идея выглядела перспективной — в неё продолжали вкладываться.

Если результаты разочаровывали — направление закрывали. Одним из проектов, который Амодеи больше не считал особенно полезным, была программа по победе в видеоигре Dota 2. Команда добилась своей цели в апреле: система OpenAI победила соперников, проект помог привлечь инвестиции Microsoft и убедил многих сотрудников в том, что стратегия масштабирования действительно работает. Но теперь, по мнению Амодеи, он выполнил свою задачу. Команду Dota 2 распустили.

Куда более перспективным Амодеи считал GPT-2. За ней стояла идея, известная как гипотеза **«чистого языка» — pure language**. Её логика была проста. Язык — главный инструмент, с помощью которого люди передают друг другу знания. Огромная часть человеческого опыта, науки, истории и культуры в конечном счёте оказывается зафиксирована в текстах. Значит, возможно, для появления AGI достаточно обучить алгоритм на колоссальном объёме языка — и больше ничего ему не понадобится. Этому подходу противостояла другая гипотеза — **grounding**, то есть «укоренение» интеллекта в реальном мире. Согласно ей, язык сам по себе недостаточен. Человеческий интеллект формируется не только благодаря словам, но и благодаря восприятию физического мира и взаимодействию с ним. Мы видим. Слышим. Двигаемся. Берём предметы в руки. Совершая действия, сталкиваемся с их последствиями. Поэтому настоящий AGI, утверждали сторонники этого подхода, сможет возникнуть лишь из сочетания языка, восприятия — например, компьютерного зрения — и способности действовать в реальной или виртуальной среде. Во внутренних документах сотрудники OpenAI подробно спорили о достоинствах обоих подходов.

И некоторые из этих дискуссий принимали довольно неприятный оборот. Пытаясь доказать центральную роль языка, исследователи в области безопасности ИИ порой проводили неудачные и оскорбительные аналогии с людьми с инвалидностью. Так становилось видно, насколько легко разговор об «интеллекте» может незаметно превратиться в опасные рассуждения о том, какие группы людей якобы умнее или менее способны, чем другие. В одном документе под заголовком «Первые аргументы в пользу центральной роли языка» было написано примерно следующее:

«Именно язык в той или иной форме отличает одичавшего человека от человека, живущего в обществе. Пример — Хелен Келлер». В комментариях на полях спор продолжался. Один исследователь написал, что слепые люди в интеллектуальном отношении ничем существенно не уступают зрячим, — значит, визуальное восприятие вовсе не обязательно для полноценного интеллекта.

Другой возразил, что слепые люди всё же находятся в заметно менее выгодном экономическом положении, и сослался на статистику занятости. Третий ответил:

«Но слепые люди всё равно гораздо способнее шимпанзе. Среди слепых есть чрезвычайно выдающиеся люди». Сам характер подобных споров показывал, насколько скользкой могла становиться попытка определить интеллект через отдельные человеческие способности.

Многие в OpenAI поначалу относились к гипотезе «чистого языка» скептически. Но GPT-2 заставила их задуматься. Ведь модель обучали всего лишь всё точнее предсказывать следующее слово.

И тем не менее вместе с этим она неожиданно становилась лучше в совершенно других задачах обработки языка. Это выглядело так, будто совершенствование одной простой способности вытягивало за собой целый набор других. Возникал соблазнительный вывод: возможно, если продолжать масштабировать GPT, увеличивать объём обучения и ещё сильнее повышать точность предсказания следующего слова, модель сама начнёт приобретать всё более широкий набор интеллектуальных возможностей. Амодеи постепенно пришёл к мысли, что масштабирование языковых моделей, пусть оно и не обязательно будет **единственным** условием появления AGI, вполне может оказаться самым быстрым путём к нему. Особенно на фоне проблем робототехнической команды. Роботизированная рука OpenAI постоянно страдала от аппаратных сбоев. Получалась худшая из возможных комбинаций: **очень дорого и очень медленно.** Языковые модели, напротив, можно было увеличивать значительно быстрее. Но здесь возникал очевидный парадокс. Если GPT-2 уже вызвала беспокойство из-за возможного распространения дезинформации, то что произойдёт, если сделать следующую модель намного мощнее?

Амодеи утверждал, что именно поэтому от масштабирования **нельзя отказываться**. И Альтман с ним соглашался. Вывод оказался почти противоположным тому, чего можно было ожидать: **OpenAI должна увеличивать языковую модель как можно быстрее — просто не выпускать её сразу в открытый доступ.** GPT-2 показала, насколько легко другие участники рынка могут получить похожие технологии. Более того, ещё до того, как OpenAI опубликовала полную GPT-2, двое аспирантов уже создали её открытую альтернативу. Значит, считал Амодеи, вопрос был не в том, будут ли другие масштабировать языковые модели. Они обязательно будут. Вопрос лишь в том, **кто сделает это первым**. Отсюда следовала стратегия. OpenAI должна была вырваться далеко вперёд, получить временное преимущество и использовать его, чтобы заранее разобраться, как сделать более мощные модели безопаснее. А когда наступит момент их публичного выпуска, более тщательно настроенная система OpenAI сможет задать стандарты безопасности для всей индустрии — подобно тому как история с задержкой публикации GPT-2 уже начала менять нормы раскрытия исследований. Кроме того, к этому времени уже появились основания считать, что опасности, связанные именно с языковыми моделями, не настолько велики, как опасались первоначально.

По крайней мере OpenAI не знала о случаях, когда GPT-2 использовали бы в крупных скоординированных кампаниях массовой дезинформации. И даже если такие угрозы были реальны, в глазах Амодеи они выглядели куда менее страшными, чем возможные экзистенциальные риски будущего AGI. Он объяснял:

«Разумеется, злоупотребления — это плохо. Меня очень тревожит возможность использования языковых моделей как оружия для дезинформации. Это действительно страшно. Но всё же это относительно конкретная и чётко очерченная проблема».

Иными словами, Амодеи видел в масштабировании GPT-2 сразу два преимущества. Во-первых, это мог быть самый быстрый путь к AGI. Во-вторых, возникающие по дороге опасности казались ему относительно управляемыми. Дезинформация. Манипуляции. Злоупотребления. Да, серьёзно. Но всё же не конец человечества. Языковые модели могли стать своеобразным полигоном: достаточно мощным, чтобы на них уже проявлялись будущие проблемы AGI, но пока не настолько мощным, чтобы последствия ошибки стали катастрофическими. Амодеи формулировал это так:

«Что такое AGI? Как он вообще будет выглядеть? Мы находимся в странном положении: мы этого не знаем. Мы не знаем и когда он появится. Поэтому мы ищем системы, которые ещё не AGI, но уже создают хотя бы некоторые из тех возможностей и трудностей, которые возникнут с AGI. И надеемся, что, если научимся правильно справляться с ними, будем готовы перейти в высшую лигу». Но вся эта логика держалась на одном чрезвычайно важном предположении: **AGI всё равно неизбежен.** Он ещё не существовал.

Никто не мог точно сказать, как он должен выглядеть. Никто не знал, когда он появится. Но внутри OpenAI его приход всё чаще воспринимался почти как неотвратимое событие. В последующие годы компания снова и снова будет оправдывать свои решения вариациями одного и того же аргумента. Раз AGI всё равно приближается, OpenAI обязана строить всё более мощные модели, чтобы подготовиться самой и подготовить общество. Даже если эти промежуточные модели создают собственные риски, их можно считать допустимой ценой за опыт, который поможет однажды предотвратить гораздо более серьёзную катастрофу.

Автор книги, однако, подвергает эту логику сомнению. Когда в начале 2023 года ChatGPT стремительно распространился по миру, один китайский исследователь ИИ предложил ей другое объяснение. По его мнению, то, что сделала OpenAI, вообще могло произойти только в Кремниевой долине. В Китае — несмотря на огромный научный потенциал страны в сфере ИИ — ни одна группа исследователей не получила бы миллиард долларов, не говоря уже о десятикратно большей сумме, для создания невероятно дорогой технологии без ясного объяснения, что именно она будет собой представлять и для чего нужна. Лишь после появления ChatGPT, когда стало видно, что огромные модели имеют реальное коммерческое применение, китайские компании и инвесторы начали вкладывать в их развитие большие средства. Но в ходе работы над книгой автор пришла к ещё более неожиданному выводу.

Даже в самой Кремниевой долине другие компании и инвесторы не бросились безоговорочно вкладывать колоссальные средства в масштабирование до появления ChatGPT. Не делали этого в таком виде даже Google и DeepMind — первоначальные главные соперники OpenAI. То, что произошло, стало возможным благодаря совершенно особому сочетанию факторов именно внутри OpenAI: её происхождению из среды миллиардеров; необычной идеологии; амбициям Сэма Альтмана; его связям; и исключительной способности привлекать деньги. Один бывший сотрудник сказал об Альтмане: **«У меня такое ощущение, что Сэм — самый амбициозный человек на планете».** Поэтому автор делает жёсткий вывод: всё произошедшее было не неизбежным. Наоборот. Колоссальные мировые расходы на гигантские модели глубокого обучения, последовавшая гонка компаний и государств за всё большим масштабом — всё это могло возникнуть именно в той уникальной среде, где и возникло. **Неизбежной была не гонка. Её кто-то начал.**

К апрельской демонстрации для Билла Гейтса в 2019 году OpenAI уже успела немного увеличить GPT-2. Но Амодеи не интересовало «немного». Если задача заключалась в том, чтобы максимально увеличить технологический отрыв OpenAI, следующая модель — GPT-3 — должна была быть **настолько большой, насколько это вообще возможно**. В рамках миллиардной инвестиции Microsoft собиралась предоставить OpenAI новый суперкомпьютер с **десятью тысячами Nvidia V100** — на тот момент одними из самых мощных графических процессоров для обучения систем глубокого обучения.

Амодеи хотел задействовать все десять тысяч одновременно. Для многих сотрудников идея казалась почти абсурдной.

До этого модель уже считалась крупной, если её обучали на нескольких десятках GPU. В ведущих университетских лабораториях MIT или Стэнфорда аспирант мог считать себя счастливчиком, получив доступ к десяти процессорам. А в университетах за пределами США — например, в Индии — несколько исследователей порой вынуждены были делить между собой один GPU. На этом фоне десять тысяч процессоров выглядели почти фантастикой. Даже многие исследователи OpenAI сомневались, что всё это вообще заработает.

Некоторые предлагали более осторожный путь: масштабировать модель постепенно, шаг за шагом, чтобы процесс оставался научно контролируемым и предсказуемым. Амодеи был непреклонен. Его поддерживали и другие руководители. Суцкевер хотел довести до конца собственную гипотезу о масштабировании трансформеров. Брокман видел возможность ещё сильнее повысить известность компании. Альтман хотел сделать максимально крупную ставку. Вскоре после этого Амодеи повысили до вице-президента по исследованиям. При этом Альтман видел ситуацию ещё и с другой стороны. Microsoft вложила **один миллиард долларов**.

А миллиард долларов создаёт ожидания соответствующего масштаба. OpenAI должна была показать нечто, что оправдает эти инвестиции. Для Амодеи огромная языковая модель была прежде всего инструментом исследований безопасности. Для Альтмана она одновременно становилась способом выполнить обещание перед Microsoft. И вскоре эти два человека начнут спорить о том, **как именно и когда выпускать GPT-3**.

Победит Альтман. Модель выведут на рынок быстрее, чем хотел Амодеи. Автор книги считает, что эти два решения — **резко увеличить масштаб GPT-3** и **как можно быстрее выпустить её в мир** — изменили всю последующую историю искусственного интеллекта. Они ускорили развитие отрасли. Запустили ожесточённую конкуренцию между компаниями и государствами. Подтолкнули дальнейший рост капитализма наблюдения и эксплуатации труда. Из-за колоссальной стоимости вычислительной инфраструктуры разработка передовых моделей начала концентрироваться в руках немногих гигантских организаций. Университеты всё меньше могли соперничать с индустрией. Аспиранты и профессора уходили в технологические компании. Независимые академические исследования слабели. Вместе с ними ослабевали и независимые механизмы контроля. Одновременно резко выросло воздействие ИИ на окружающую среду — причём из-за отсутствия прозрачности и достаточного регулирования даже внешние эксперты и правительства долгое время не могли точно оценить масштаб этих последствий. Но осенью 2019 года всё это ещё оставалось в будущем. Тогда Амодеи сформировал команду, получившую название **Nest**. В неё вошли преимущественно специалисты по безопасности ИИ. Их задача состояла в том, чтобы держать разработку GPT-3 под жёстким внутренним контролем. После этого началась настоящая гонка за масштабом.

По сути, GPT-3 была всё той же GPT-2. Только ей дали **невообразимо больше данных и вычислительных ресурсов**. Масштаб вырос настолько, что результат начал восприниматься уже не просто как количественное улучшение, а почти как качественный скачок.

Но использование десяти тысяч процессоров породило совершенно новые технические проблемы. Каждый отдельный GPU мог дать сбой во время обучения — примерно как обычный ноутбук может зависнуть, если на нём одновременно запущено слишком много программ. Проблема состояла в том, что если ломался **один** процессор, могла остановиться вся система. А тогда обучение пришлось бы начинать заново. На одном GPU вероятность сбоя невелика. Но когда их десять тысяч, вероятность того, что хотя бы один выйдет из строя, резко возрастает. А команда Nest ожидала, что обучение будет продолжаться несколько месяцев. Ошибка спустя недели работы стоила бы огромных денег и огромного количества времени.

Поэтому инженерам пришлось придумать способ сохранять состояние обучения так, чтобы после сбоя система могла продолжить **точно с того места, где остановилась**. Была и другая проблема. Как распределить гигантскую модель между десятью тысячами процессоров? Нужно ли разбить её на

десятки частей? На сотни? На тысячи? Каждую часть обучать на отдельной группе GPU, а потом объединять результаты?

Этот процесс называется **sharding — шардирование**. Ещё сложнее оказался вопрос данных. Чтобы сохранять высокое качество модели при росте числа параметров и вычислений, нужно было соответствующим образом увеличивать и объём обучающего материала. Если параметров слишком много, а данных недостаточно, модель начинает не столько учиться закономерностям, сколько **запоминать** обучающие тексты и воспроизводить их почти слово в слово. Для GPT-2 Рэдфорд достаточно тщательно отбирал материал. Он собирал статьи и веб-страницы, ссылки на которые публиковались на Reddit и получали не менее трёх положительных голосов. Получился набор WebText: около **40 гигабайт**; примерно **восемь миллионов документов**. Для GPT-2 этого хватало.

Для GPT-3 — уже нет. Команда Nest расширила корпус. Добавили значительно больше ссылок с Reddit. Добавили англоязычную Википедию. А также загадочный набор данных под названием **Books2**. OpenAI никогда подробно не раскрывала его происхождение. Однако два человека, знакомые с этим набором, рассказали автору книги, что там находились опубликованные книги, полученные из Library Genesis — теневого интернет-хранилища пиратских книг и научных статей. Позднее это станет юридической проблемой. В 2023 году Authors Guild и семнадцать писателей, среди которых Джордж Р. Р. Мартин и Джоди Пиколт, подадут в суд на OpenAI и Microsoft, обвинив их в массовом нарушении авторских прав. В марте 2024 года OpenAI заявит, что эти наборы данных были удалены и после GPT-3.5 больше не использовались для обучения. Но даже этого оказалось мало.

Тогда команда обратилась к **Common Crawl** — гигантскому общедоступному архиву интернета. Речь шла уже о петабайтах данных — миллионах гигабайт текста, автоматически собранного со всего веба. Рэдфорд намеренно избегал Common Crawl при создании GPT-2. Причина была проста: качество материала там было слишком низким. Но теперь требовался масштаб. Чтобы хоть как-то отделить полезный текст от мусора, команда Nest обучила отдельную модель искать в Common Crawl материалы, похожие на статьи Википедии. Логика была приблизительно такой: если текст выглядит как Википедия, есть шанс, что его качество тоже будет ближе к Википедии. В корпус добавили и некоторое количество текстов не на английском языке, однако в итоге они составили всего около **7 процентов** данных. Даже после фильтрации Common Crawl использовался с меньшим весом, чем более качественные источники. Получалась любопытная картина: **GPT-2 стала пиком качества обучающих данных. После неё качество начало снижаться.**

Когда спустя два года пришло время собирать данные для GPT-4, давление количества стало ещё сильнее. Фильтрацию Common Crawl в значительной степени убрали, и в набор хлынуло гораздо больше сырых интернет-данных. Благодаря партнёрству с Microsoft OpenAI также получила полный массив GitHub — принадлежащего Microsoft крупнейшего онлайн-хранилища программного кода. Но и этого оказалось недостаточно. Сотрудники собирали всё, до чего могли дотянуться в интернете: ссылки из Twitter; расшифровки YouTube-видео; нишевые блоги; готовые интернет-архивы; тексты с Pastebin; и множество других источников. Автор описывает действовавший подход довольно просто: **если на сайте не было прямого запрета на автоматический сбор данных, его содержимое считалось доступным.**

Некоторые исследователи Google с раздражением наблюдали за тем, насколько свободно OpenAI идёт на юридические риски. В Google правила использования данных были гораздо строже. Компания придерживалась сложных процедур соответствия законодательству, включая европейский GDPR. Парадоксально, но из-за этого OpenAI порой получала более свободный доступ даже к данным, принадлежащим самому Google, чем исследователи Google. Например, сотрудники OpenAI могли собирать и расшифровывать видео с YouTube. А исследователям внутри Google приходилось проходить многочисленные согласования, чтобы не нарушить условия лицензирования пользовательского контента YouTube. OpenAI была гораздо меньше связана такими ограничениями. Автор связывает это с классической культурой технологических стартапов Кремниевой долины: **сначала зайти в юридически серую зону, а уже потом разбираться с последствиями.** Подобным образом в своё время действовали Airbnb, Uber или Coinbase, бросая вызов правилам традиционных отраслей.

Но снижение требований к качеству данных имело ещё одно, гораздо менее заметное последствие. Оно изменило сам характер человеческого труда, необходимого для создания ИИ. Технологическая индустрия и раньше использовала дешёвый труд людей из экономически уязвимых регионов для подготовки данных: они классифицировали тексты; размечали изображения; сортировали материалы. Но после GPT-3, когда гигантские низкокачественные массивы стали нормой, изменилась и сама работа. Теперь людям всё чаще приходилось иметь дело с отвратительным и психологически тяжёлым контентом. Насилие. Сексуальное насилие. Самоповреждения. Ненависть. Похожая проблема раньше возникла в социальных сетях. Генеративная модель обучалась фактически на огромном срезе человеческого интернета, а значит могла воспроизводить его самые мерзкие части для сотен миллионов пользователей.

Следовательно, если нельзя было полностью очистить **входные данные**, приходилось гораздо жёстче контролировать **выход модели**. Райан Коллн, руководитель платформы Appen, связывающей технологические компании с работниками по обработке данных, описывал это как смену целой парадигмы. В традиционном машинном обучении, объяснял он, выход системы контролировали прежде всего через вход: модель учится только на тех примерах, которые ей дают. Но у генеративного ИИ входом становится почти **весь корпус человеческой культуры**. Поэтому контролировать приходится уже не только то, чем модель кормят, но и то, что она выдаёт наружу.

В 2023 году исследовательница Абеба Бирхане и её коллеги предложили выражение **«законы масштабирования ненависти» — hate scaling laws**. Так они критиковали саму идею обучать огромные системы на почти неотфильтрованных «болотах данных». Они изучили два открытых набора изображений и подписей — LAION-400M и LAION-2B-en, содержащие соответственно около 400 миллионов и 2 миллиардов изображений. Оба были собраны на основе Common Crawl. Исследователи обнаружили: по мере роста размера набора увеличивалось и количество ненавистнического, оскорбительного и дискриминационного материала. А модели, обученные на нём, воспроизводили эти искажения. Например, система, обученная на наборе из двух миллиардов изображений, значительно чаще маркировала лица чернокожих мужчин как связанные с преступностью, чем модель, обученная на меньшем наборе. Позже в том же году исследование Стэнфорда другого огромного набора — LAION-5B — обнаружило в нём тысячи изображений, которые были подтверждены или подозревались как материалы сексуального насилия над детьми. Чтобы контролировать выходы своих моделей, OpenAI стала активно использовать человеческую модерацию. Компания нанимала работников в Кении — в среднем менее чем за два доллара в час — для создания автоматизированного фильтра нежелательного контента. Об этом впервые подробно сообщил журналист *Time* Билли Перриго.

Кроме того, более тысячи подрядчиков по всему миру привлекались к работе по **RLHF — обучению с подкреплением на основе человеческой обратной связи**. То самое решение, которое когда-то помогло виртуальной Т-образной фигурке научиться делать сальто, теперь применяли к гигантским языковым моделям. Люди снова и снова задавали модели вопросы. Сравнивали ответы. Оценивали их. Показывали системе, какой ответ считается лучшим, приемлемым или безопасным. Так исследователи пытались приручить всё более мощную машину. Немецкая художница и режиссёр Хито Штайерль, снявшая документальный фильм о сирийских беженцах, работающих в сфере обработки данных, сформулировала проблему ещё шире.

По её мнению, психологически разрушительный контент закономерно накапливается там, где в основе сбора данных лежит массовое наблюдение и бесконтрольное поглощение интернет-информации. И если мы действительно хотим решить проблему, недостаточно лишь строить всё лучшие фильтры. Нужно вернуться к исходному вопросу: **что именно содержится в этих данных?** И ещё глубже: **почему мы вообще считаем нормальным собирать их в таких масштабах — без разбора и почти без ограничений?**

ГЛАВА 6
ВОСХОЖДЕНИЕ

Ещё в начале своей карьеры Альтман заметил одну вещь: новый генеральный директор добивается успеха только тогда, когда фактически **заново основывает компанию**. Когда он возглавил Y Combinator, именно так и поступил. Он создал новые программы, в том числе новый инвестиционный фонд, чтобы YC мог поддерживать стартапы на разных стадиях развития. Он начал продвигать так называемые hard tech — сложные технологические проекты, требующие серьёзных научных прорывов: ядерный синтез, квантовые вычисления, беспилотные автомобили. Y Combinator и до этого был престижным брендом, но при Альтмане его влияние выросло колоссально: вместо нескольких сотен компаний он начал охватывать тысячи стартапов в год. YC превратился в один из главных центров притяжения Кремниевой долины. После ухода с поста президента Альтман говорил: «Больше всего я горжусь тем, что мы действительно построили империю».

С завершением эпохи YC для него началась эпоха OpenAI. В марте 2019 года, когда Альтман полностью переключился на OpenAI, он принёс туда тот же агрессивный стиль. Он не хотел, чтобы OpenAI была **одной из ведущих организаций в мире искусственного интеллекта**. Он хотел, чтобы она стала **единственной по-настоящему главной**. Годами через YC и другие площадки Альтман учил основателей стартапов смотреть на рынок как на игру: **победитель получает всё.** Если хочешь выжить, говорил он, нужно двигаться быстро и безжалостно, обгонять соперников — и продолжать держать их позади.

Особое значение для Альтмана имело число **десять**. Эта идея пришла от Питера Тиля и его стратегии построения монополий. На лекции 2014 года в Стэнфорде Тиль говорил: «Моё немного безумное и довольно условное правило таково: ваша технология должна быть примерно в десять раз лучше ближайшей альтернативы». Amazon, например, научилась продавать в десять раз больше книг, чем обычные книжные магазины. PayPal научился переводить деньги примерно в десять раз быстрее, чем традиционные банковские чеки.

Тиль объяснял: нужно иметь настолько мощное преимущество в каком-то ключевом параметре, чтобы конкуренты просто не могли легко вас догнать. Для Альтмана «в десять раз» превратилось почти в универсальную формулу. Стартап должен не просто выйти на рынок с технологией, которая в десять раз лучше. Каждое новое поколение продукта тоже должно быть в десять раз сильнее предыдущего.

И крайне важна скорость циклов. В 2017 году Альтман говорил молодым предпринимателям: — Если вы обновляете продукт каждую неделю, а конкурент — раз в три месяца, вы просто оставите его далеко позади. В YC он постоянно подталкивал партнёров увеличивать число финансируемых компаний в десять раз. Потом — ещё в десять. Потом — снова.

Альтман однажды сказал: «Когда-нибудь мы будем финансировать все компании в мире». Его преемник в YC Джефф Ралстон вспоминал: — Сэм был первым человеком, от которого я услышал мысль, что благодаря работе первых основателей и бренду YC мы фактически уже стали монополией в этой сфере.

В OpenAI Альтман намеревался использовать ту же модель. В конце 2019 года он разослал сотрудникам меморандум, где изложил долгосрочное видение компании. К концу 2020 года OpenAI, по его замыслу, должна была стать **номером один** сразу по четырём направлениям: технические результаты; вычислительные ресурсы; деньги — чтобы покупать ещё больше вычислительных ресурсов; и подготовленность — то есть безопасность, защита и способность организации выдерживать ситуации экстремального давления. Но главным был первый пункт. Если OpenAI действительно хотела выполнить свою миссию, она должна была либо первой создать полезный AGI, либо настолько доминировать в этой области, чтобы всё равно определять, каким будет его развитие. Альтман писал: «Теоретически мы могли бы замедлить работу над возможностями ИИ. Но с учётом

того, как быстро продвигаются другие, нам, скорее всего, придётся двигаться очень быстро, если мы хотим иметь серьёзное влияние на AGI».

И чем больше конкурентов поймут стратегию OpenAI, тем сильнее станет давление. Позже в том же документе он добавлял: «До AGI нам потребуется ещё много скачков в десять раз. Мы всегда должны стремиться к драматическим результатам, а не к постепенным улучшениям».

Одним из важнейших условий успеха было сохранить хорошие отношения с Microsoft. Именно Microsoft обеспечивала OpenAI вычислительное преимущество. Если сотрудничество окажется успешным, компания уже дала понять, что вложит гораздо больше одного миллиарда долларов. Альтман писал: «Мы хотели бы, чтобы Microsoft оставалась нашим главным партнёром на всём пути». И дальше: «По меньшей мере следующие пять лет они способны давать нам самые мощные суперкомпьютеры в мире». Но это означало перемены. OpenAI уже не могла оставаться свободной академической лабораторией, где люди просто исследуют всё, что кажется интересным.

Теперь нужно было сосредоточиться и на коммерциализации. Нужно было давать Microsoft конкретную пользу. А уже заработанные деньги и ресурсы позволят финансировать другие исследования. Альтман перефразировал известную фразу Уолта Диснея: «Нам нужно зарабатывать больше денег, чтобы делать больше исследований, а не делать больше исследований, чтобы зарабатывать больше денег».

Одновременно компания должна была стать **менее прозрачной**. Альтман предупреждал: чем ближе OpenAI будет подходить к AGI, тем опаснее станет открыто рассказывать о своей работе. Он использовал термин **infohazard** — информационная опасность: знание само по себе может стать источником риска. Следовало ограничить публикацию исследований. Строже контролировать выпуск моделей.

Ужесточить конфиденциальность. И публично показывать только отдельные способности систем, а не общий уровень прогресса. При этом каждый сотрудник должен был жить так, будто: «каждое наше решение и каждый разговор однажды будут расследованы и окажутся на первой полосе The New York Times».

Но полная тишина тоже не годилась. Мир, писал Альтман, всё равно должен думать, что OpenAI **в чём-то побеждает**. Это заставит влиятельных людей охотнее помогать компании. А политики самого высокого уровня — президенты и их представители — будут приходить в OpenAI за советом, когда потребуется принимать важные решения. Поэтому Альтман предлагал: **как минимум раз в год показывать миру что-нибудь действительно впечатляющее.**

И наконец, компании требовалось стать серьёзнее и сплочённее. В меморандуме Альтман привёл цитату адмирала ВМС США Хаймана Риковера — «отца атомного флота», сыгравшего ключевую роль в создании первых атомных подводных лодок.

Эти слова в первые годы OpenAI даже были написаны на стене офиса:

«Я считаю, что каждый из нас обязан действовать так, словно судьба мира зависит от него. Конечно, один человек не может сделать всё. Но один человек способен изменить многое. Каждый из нас обязан приложить свои способности к решению широкого круга человеческих проблем. Именно с этим убеждением мы должны прямо смотреть в лицо своей ответственности перед будущим. Мы должны жить ради будущего человеческого рода, а не ради собственного комфорта или успеха». Под этой цитатой Альтман написал: «Создание AGI, который принесёт пользу человечеству, возможно, самый важный проект в мире». И дальше: «Мы обязаны ставить миссию выше любых личных предпочтений». А потом добавил почти бытовое правило: «Мелочи должны оставаться без драм, чтобы мы могли сохранять способность к настоящей драме для действительно важных вещей — а таких будет много».

Драма, впрочем, уже начиналась. Мелкие трещины внутри компании постепенно превращались в настоящие разломы. Брокман и Дарио с Даниэлой Амодеи, которые когда-то легко называли друг

друга друзьями, всё чаще конфликтовали. Одной из причин стала Dota 2. Амодеи всё меньше считал этот проект приоритетным. Брокмана это раздражало: ему казалось, что его вклад недооценивают. А затем возникла ещё более болезненная тема — **вычислительные ресурсы.** Когда-то именно Dota 2 была самым тяжёлым в вычислительном отношении проектом OpenAI. Теперь Амодеи сосредоточил огромную долю вычислительной мощности на команде Nest, которая работала над GPT-3. Брокман воспринимал это тяжело. А Амодеи и его люди, в свою очередь, считали Брокмана трудным партнёром и не хотели пускать его в разработку языковых моделей.

Конфликт руководителей начал распространяться на их сторонников. Во время работы над Dota 2 Брокман чрезвычайно сблизился с частью своей команды. Их сплотили длинные рабочие дни, сильнейший стресс и даже спонтанная поездка на Гавайи. Особенно близкими к нему стали Якуб Пахоцкий и Шимон Сидор — польские учёные, лучшие друзья и соседи по комнате. У Амодеи сформировалась другая группа. Его команды по безопасности ИИ, а особенно ядро Nest, объединяло беспокойство по поводу неконтролируемого ИИ и возможных экзистенциальных рисков. Эти люди всё больше закрывались от остальных. Создавали приватные Slack-каналы. Вели документы, недоступные даже некоторым руководителям. Других сотрудников это раздражало. Они видели, как у них забирают вычислительные ресурсы. И одновременно всё меньше понимали, что происходит в главном исследовательском направлении компании.

Но люди Амодеи начали тревожиться и из-за поведения самого Альтмана. Вскоре после заключения сделки с Microsoft некоторые из них с удивлением узнали, насколько далеко Альтман зашёл в обещаниях о том, к каким технологиям Microsoft получит доступ в обмен на инвестиции. Условия выглядели иначе, чем они поняли из разговоров с Альтманом. И возник тревожный вопрос: что, если в моделях действительно обнаружатся серьёзные проблемы безопасности? Смогут ли они тогда остановить выпуск системы?

Или юридические обязательства перед Microsoft уже сделают это практически невозможным? У группы начали появляться серьёзные сомнения в честности Альтмана. Один сотрудник объяснял: — Мы все прагматики. Понятно, что мы привлекаем деньги и будем заниматься коммерцией. Если ты заключаешь сделки постоянно, как Сэм, всё может выглядеть совершенно нормально: «Давайте обменяем одно на другое, потом ещё на что-то». Но я смотрю и думаю: «Мы сейчас обмениваем вещь, которую сами ещё не до конца понимаем». И мне очень неуютно от того, куда это нас привязывает.

На всё это накладывалась растущая паранойя внутри OpenAI. В группе безопасности она прежде всего была связана с убеждением: появляется всё больше признаков того, что мощный ИИ с неправильными целями действительно может привести к катастрофе. Один эпизод особенно запомнился сотрудникам. В 2019 году исследователи экспериментировали с моделью, появившейся после GPT-2 и имевшей примерно вдвое больше параметров. Они тестировали **обучение с подкреплением на основе человеческой обратной связи — RLHF**. Цель была простой: научить модель генерировать более дружелюбный и позитивный контент и избегать оскорбительного.

Однажды поздно вечером исследователь изменил код. И перед уходом оставил процесс работать всю ночь. В коде была одна опечатка. Всего один знак. Там, где должен был стоять минус, стоял плюс. Но из-за этого RLHF заработал **в обратную сторону**. Вместо того чтобы уменьшать количество непристойного контента, модель начала специально стремиться к нему. Утром исследователи обнаружили, что GPT-2 буквально на любой запрос отвечает крайне грубым, непристойным и сексуально откровенным текстом. Ситуация была одновременно очень смешной и немного пугающей. Исправив ошибку, исследователь оставил комментарий в коде: **«Давайте не будем создавать минимизатор полезности».** Шутка отсылала к классическому страху: неправильно сформулированная цель может заставить систему делать совсем не то, что предполагал человек.

Одновременно сотрудники всё чаще думали о другом. Что будет, если конкуренты поймут главный секрет OpenAI? А секрет казался почти смешным своей простотой. Люди шутили: «Секрет

того, как работает наша система, можно написать на рисовом зёрнышке». И это слово было: **scale — масштабирование.** Больше данных.

Больше параметров. Больше вычислений. По той же причине многие боялись, что слишком мощные системы попадут к опасным людям. Руководство постоянно говорило о Китае, России и Северной Корее. И подчёркивало: AGI должен оставаться под контролем американской организации. Некоторых иностранных сотрудников это раздражало. За обедом они спрашивали: — Почему обязательно американской? Почему не европейской? Почему не китайской?

Такие разговоры всё чаще возвращались к старому сравнению Альтмана: **OpenAI как Манхэттенский проект.** И у сотрудников возникал странный вопрос:

что же они вообще строят? Эквивалент атомной бомбы? Это резко контрастировало с ещё недавней культурой молодой академической лаборатории. По пятницам люди расслаблялись после тяжёлой недели. Пили вино. Слушали музыку. Кто-нибудь из коллег садился за офисное пианино и играл до поздней ночи. Но теперь настроение менялось. Любая случайность могла вызвать тревогу. Однажды журналист прошёл за сотрудником через ворота закрытой парковки и таким образом проник в здание. В другой раз нашли неизвестную USB-флешку. Сразу возникло подозрение: а вдруг на ней вредоносное ПО?

Может быть, это попытка взлома? Флешку проверили на полностью изолированном от интернета компьютере. Там ничего опасного не оказалось. Но страх уже был симптомом новой атмосферы. Амодеи как минимум дважды писал важные стратегические документы на изолированном компьютере. Компьютер подключали напрямую к принтеру. Документы распространяли только на бумаге. Он серьёзно боялся, что государственные спецслужбы могут украсть секреты OpenAI и использовать их для создания собственного мощного ИИ. Один сотрудник вспоминал: — Никто не был готов к такой ответственности. Люди ночами не спали.

Сам Альтман тоже боялся утечек. Его беспокоили сотрудники Neuralink, с которыми OpenAI всё ещё делила офисное пространство после ухода Маска. Он беспокоился и о самом Маске. У того была мощная система личной безопасности: водители, телохранители, специалисты. Альтман хорошо понимал, что возможности Маска несопоставимы с его собственными. В какой-то момент он тайно заказал проверку офиса на электронную слежку — искали возможные скрытые устройства, которые Маск мог оставить для наблюдения за OpenAI.

Сотрудникам Альтман объяснял необходимость всё большей закрытости угрозой со стороны геополитических соперников США. OpenAI должна работать максимально быстро. И одновременно становиться всё менее открытой. В документе о будущем компании он писал: «Мы обязаны считать себя ответственными за хороший исход для всего мира». Но тут же добавлял: «Если авторитарное государство создаст AGI раньше нас и использует его во зло, это тоже будет означать провал нашей миссии». Вывод: **чтобы выполнить миссию, OpenAI почти наверняка должна очень быстро развивать свои технические возможности.**

После этого Альтман ещё сильнее ужесточил безопасность. Руководители спорили: на что должна быть похожа OpenAI? На крупную корпорацию, защищающую коммерческие технологии? Или на государственную структуру, охраняющую секреты национальной безопасности? Минимум, с которым все согласились: нужно максимально защищать **веса моделей**. Это ключевые данные полностью обученной нейросети. Если их украсть, можно воспроизвести саму модель. А это означало сразу две угрозы: опасные люди получат мощную систему; OpenAI потеряет конкурентное преимущество.

Поначалу в компании даже не было отдельной службы безопасности. Поэтому Альтман поручил одному из сотрудников инфраструктурной команды продумать защиту моделей. Причём не только от корпоративных шпионов или спецслужб. Но и от **собственных сотрудников**. В кибербезопасности это обычное понятие: **insider threat — внутренняя угроза**. Человек изнутри может сознательно

украсть информацию. Или его могут обмануть и заставить выдать её. В частном разговоре Альтман признал, что Суцкевер, например, теоретически может оказаться уязвим для второго варианта. Он был идеальной целью: выдающийся учёный, очень высокий уровень доступа, но не самый житейски осторожный человек.

У Суцкевера, впрочем, были свои страхи. Как знаменитый учёный в довольно закрытом и социально неловком мире исследователей ИИ, он уже сталкивался с навязчивыми поклонниками. Иногда незнакомцы пытались проникнуть в офис OpenAI просто для того, чтобы увидеть его. Как и Амодеи, Суцкевер боялся интереса со стороны недобросовестных государств. Он иногда подозревал: а вдруг люди, слишком настойчиво просящие его совета, на самом деле иностранные агенты? Он даже обсуждал с коллегами совершенно экстремальный сценарий: что делать, если кто-нибудь **отрежет ему руку**, чтобы использовать ладонь для биометрического доступа к секретам OpenAI? Суцкевер хотел нанимать меньше людей. Маленькая команда, считал он, снижает риск проникновения. Вместе с Пахоцким и Сидором он предложил построить специальный защищённый бункер. Внутри — компьютер, полностью отключённый от интернета. Там будут храниться веса моделей. Но идея была практически бессмысленной: модели сначала всё равно приходилось обучать на серверах Microsoft. Поэтому проект так и не реализовали.

Большинство сотрудников почти не замечало происходившего. Но цифровой контроль усиливался. Устанавливались корпоративные системы мониторинга. Усиливалась физическая безопасность. Ворота парковки сделали прочнее. На нескольких дверях появились специальные **коды тревоги**. Внешне человек вводил обычный код. Но секретная комбинация одновременно отправляла службе безопасности сигнал: **человека заставляют открыть дверь под угрозой.** Компания даже запросила у подрядчиков стоимость укрепления серверной так, чтобы она могла выдержать обстрел из автоматического оружия. От этой идеи позже отказались.

В своём стратегическом документе Альтман уже прямо признавал: компания раскалывается. Он писал: «У нас в OpenAI есть как минимум три клана». Условно: **Exploratory Research — исследователи; Safety — безопасность; Startup — стартаперы.** Первый клан хотел прежде всего расширять возможности ИИ. Второй — действовать максимально ответственно. Третий — двигаться быстро и добиваться результата.

У каждого была важная ценность. Исследовательский клан: «Мы будем искать важные новые идеи, даже если много раз потерпим неудачу». Клан безопасности: «Мы будем непоколебимо стремиться поступать правильно». Клан стартапа: «Мы найдём способ это сделать». Но Альтман предупреждал: «Нам нужно избежать племенной войны». Для успеха все три клана должны стать одним племенем — сохранив сильные стороны каждого — и вместе работать над AGI, который принесёт человечеству максимальную пользу. Хотя Альтман не называл имён, сотрудники прекрасно понимали, кого он имеет в виду. Суцкевер — **Exploratory Research.** Амодеи и его сторонники — **Safety.** Брокман — **Startup.** А вскоре началась пандемия. Все разошлись по домам. И теперь каждому «клану» стало ещё легче жить отдельно от остальных.

Амодеи тем временем торопил свою команду. Как и с GPT-2, они начали последовательно обучать всё более крупные модели. Цель — модель на **10 тысячах GPU** с **175 миллиардами параметров**. Версии называли в алфавитном порядке в честь знаменитых учёных. **Ada** — в честь Ады Лавлейс, которую считают первым программистом. **Babbage** — в честь Чарльза Бэббиджа, придумавшего механический прообраз цифрового компьютера. **Curie** — в честь Марии Кюри, первой женщины — лауреата Нобелевской премии и первого человека, получившего её дважды. И наконец: **Davinci** — в честь Леонардо да Винчи. Задача была двойной. Во-первых, проверить: продолжают ли законы масштабирования работать на совершенно новом уровне? Во-вторых, постепенно решить технические проблемы, возникающие на каждой ступени: оборудование, данные, обучение. Команда Nest регулярно показывала результаты всей компании. И возбуждение росло. Один исследователь

вспоминал: — Невозможно передать, насколько безумно было это видеть. Я никогда в жизни не видел ничего подобного.

Параллельно Альтман и Брокман готовили коммерческий план. В конце января 2020 года Брокман написал первые строки кода для **API GPT-3**. API должен был позволить компаниям и разработчикам пользоваться возможностями модели — но не получать её веса. То есть использовать GPT-3 внутри собственных продуктов, не обладая самой моделью. OpenAI разделилась на две части. Миру Мурати повысили до вице-президента нового подразделения **Applied** — прикладного направления. Она отвечала за API и коммерциализацию. Питер Велиндер, раньше руководивший робототехникой, перешёл в продукт. Также наняли новых руководителей продуктовых и инженерных команд. Все остальные автоматически оказались в подразделении **Research**.

Но это разделение только усилило старый раскол. В Applied начали приходить люди из других стартапов. И тем самым усиливался «клан Startup». Исследователи относились к этому настороженно: не превратится ли OpenAI просто в очередную технологическую компанию Кремниевой долины? А группа Safety была возмущена гораздо сильнее. По их мнению, выпуск GPT-3 через API слишком скоро уничтожит главное преимущество, ради которого OpenAI вообще так спешила масштабироваться: **запас времени на исследования безопасности.** Если вы первыми создаёте более мощную модель, смысл лидерства в том, что у вас есть время понять, как сделать её безопасной. А если тут же выпустить её на рынок —

весь этот запас исчезает. Applied отвечала: без реального использования невозможно понять, как модель будет вести себя в мире. Кроме того, API — самый контролируемый способ выпуска. OpenAI сама решает, кому дать доступ. Может наблюдать за использованием. Может собирать данные о злоупотреблениях. И получает деньги. На общих собраниях Альтман пытался примирить стороны: API, говорил он, поможет и коммерческому направлению, и безопасности. Доход позволит ещё больше инвестировать в исследования рисков.

Когда обучение GPT-3 завершалось, сотрудники начали активно играть с моделью. Проверяли границы её возможностей. Тестировали первые версии API. В компании устроили хакатон. Люди придумывали всё новые применения. Но каждый новый прототип только усиливал конфликт. Applied и многие исследователи смотрели на демонстрации с восторгом. Safety видела в них доказательство обратного: модель слишком мощная, чтобы выпускать её без гораздо более глубокого тестирования.

Особенно спорной оказалась неожиданная способность GPT-3 — **генерировать программный код.** Команда Nest специально этому модель не учила. Но данные для обучения собирались в том числе через ссылки Reddit и Common Crawl. В них попадались фрагменты программ, которые инженеры публиковали на форумах, задавая вопросы или делясь советами. Поэтому модель случайно получила немалое количество кода. И начала неплохо работать с языками программирования. Для исследователей и Applied это было потрясающе. Во-первых, технический прорыв. Во-вторых, инструмент, способный ускорить сами исследования ИИ. В-третьих, отличный продукт. Но в Safety испугались. Если ИИ умеет писать код — что, если однажды он сможет использовать этот навык, чтобы изменять **самого себя**? Тогда развитие может ускориться. Контроль человека — ослабнуть. А риск крайне опасных сценариев — вырасти.

Суцкевер и Войцех Заремба создали отдельную команду, которая должна была специально разработать модель для генерации кода. Но на первой же встрече выяснилось: Амодеи уже занимается собственным аналогичным проектом. И объединяться не хочет. При всей своей тревоге он исходил из той же логики, что и раньше: лучший способ сделать опасную технологию безопасной — **создать её раньше всех.** Даже раньше других команд внутри OpenAI. Тогда у вас будет больше времени на исследование рисков. В итоге две группы внутри одной компании параллельно делали почти одно и

то же. Один исследователь вспоминал: — Со стороны всё это выглядело каким-то безумием в духе «Игры престолов».

Спор о выпуске GPT-3 через API тянулся до конца весны. Safety настаивала: отложить запуск максимально надолго. Applied готовилась запускать API и утверждала: модель должна столкнуться с реальным миром, иначе её невозможно улучшить. Одновременно появилась и третья группа обеспокоенных сотрудников. Их волновали не столько далёкие экзистенциальные риски, сколько совершенно реальные события в США. К маю 2020 года пандемия вызвала рост безработицы быстрее, чем во время Великой рецессии. В том же месяце полицейский Дерек Шовин убил в Миннеаполисе сорокашестилетнего Джорджа Флойда. По США и всему миру прокатились массовые протесты Black Lives Matter. Приближались президентские выборы. И на этом фоне выпуск системы, способной массово генерировать убедительные тексты, казался особенно рискованным.

Но затем внутри OpenAI распространился слух: **Google вот-вот выпустит собственную крупную языковую модель.** Это звучало правдоподобно. В начале года Google уже рассказала о чат-боте Meena. В его основе была модель примерно в 1,7 раза крупнее GPT-2. Значит, рассуждали в OpenAI, Google вполне может сейчас строить систему масштаба GPT-3. Этот слух фактически решил судьбу запуска. Если столь же мощная модель всё равно скоро появится, аргумент Safety «лучше пока ничего не выпускать» становился слабее. В июне OpenAI объявила API и открыла форму заявок на ранний доступ. Приоритет отдавали крупным компаниям, которые Applied считала достаточно ответственными. Параллельно существовала большая внутренняя таблица. Сотрудники могли вписывать туда людей, которых хотели пропустить без очереди: родственников, друзей, любимых знаменитостей.

Модель Google, которой все так боялись, в итоге **не появилась**. Google действительно работала над более крупной системой — LaMDA. Но она всё равно была немного меньше GPT-3. И компания решила не выпускать её до появления ChatGPT. Руководители Google сочли, что LaMDA не соответствует их стандартам этичного ИИ. Некоторые сотрудники вспоминали и старый скандал Microsoft. В 2016 году компания выпустила чат-бота Tay. Пользователи быстро заставили его повторять расистские, женоненавистнические и пронацистские высказывания. Так GPT-3 стал не последним случаем, когда OpenAI торопилась с выпуском технологии из-за **преувеличенного страха перед конкурентами**.

Для людей внутри мира ИИ GPT-3 стала примерно тем же, чем ChatGPT позднее станет для широкой публики. ChatGPT в 2022 году добавит: удобный веб-интерфейс, диалоговый режим, дополнительные меры безопасности, бесплатный доступ. Но многие ключевые способности, которыми через два года будет потрясён весь мир, разработчики уже увидели через API GPT-3 в 2020-м. И реакция была похожей: **неверие и восторг.** GPT-3 превосходила GPT-2 практически во всём. Никто до этого не видел системы, способной с одинаковой лёгкостью генерировать: эссе, сценарии, программный код. Уже одна эта универсальность была техническим чудом. Старые языковые модели обычно умели делать одну задачу, для которой их специально обучили. GPT-3 демонстрировала ещё одну давно желанную способность: **быстрое обобщение.** Достаточно было показать ей несколько примеров новой задачи — и она уже начинала её выполнять.

На конференции NeurIPS того года статья OpenAI о GPT-3 получила одну из главных научных наград. Даже сотрудники были удивлены. Теперь статус лаборатории как одного из мировых лидеров стал практически неоспоримым. И именно то, что предсказывало руководство, начало происходить. Сильный бренд облегчал найм. Людей стало проще удерживать. А деньги OpenAI LP наконец позволили конкурировать с Google и DeepMind по зарплатам.

В октябре 2020 года Альтман нанял Стива Даулинга — опытного специалиста, который ранее руководил коммуникациями Apple. Он стал вице-президентом OpenAI по коммуникациям. Ему же поручили отношения с государством. Альтман считал крайне важным: заранее объяснять политикам, на что будет способен ИИ.

Позднее, после ухода Джека Кларка, Даулинг пригласит Анну Маканджу — уважаемого бывшего советника администрации Обамы, также работавшую в Facebook и Starlink. Она возглавит политику и международные отношения OpenAI.

Applied тем временем хотела воспользоваться успехом GPT-3 и ещё быстрее развивать коммерциализацию. Но Safety сопротивлялась почти на каждом шагу. Для людей Амодеи сама поспешность запуска уже была ошибкой.

Теперь, считали они, нужно не распространять GPT-3 ещё шире, а срочно устранить её недостатки. В API тогда вообще не было нормальной фильтрации контента. Ответы модели ещё не были доработаны RLHF. На встречах обе стороны бесконечно пытались найти компромисс. Но вместо этого разговаривали словно на разных языках. Питер Велиндер однажды раздражённо заметил: каждая такая встреча напоминает ему секретное руководство американских спецслужб 1944 года по **ненасильственному саботажу**, рассекреченное в 2008-м. Там среди советов по снижению эффективности организации были такие: Говорите как можно чаще и как можно дольше. Как можно чаще поднимайте не относящиеся к делу вопросы. Спорьте о точных формулировках писем, протоколов и решений. Возвращайтесь к вопросам, которые уже были решены на прошлой встрече. Задавайте бесконечное количество вопросов. По мнению Велиндера, примерно так и выглядело внутреннее общение OpenAI.

Неприязнь давно вышла за пределы совещаний. Для людей из Applied почти любой цифровой канал превращался в поле боя. Стоило продуктовому сотруднику что-нибудь написать в Slack —

появлялись десятки обеспокоенных ответов от Safety. Мурати или Велиндер могли создать Google-документ с новой идеей коммерческой стратегии —

и через некоторое время он был буквально весь покрыт жёлтыми комментариями. Люди из Applied смотрели на происходящее и думали: GPT-3 уже вышла. Мир не закончился.

Значит, Safety просто истерит и живёт в какой-то другой реальности. Safety воспринимала это совершенно иначе. Для них речь шла о принципах. О прецеденте. OpenAI должна была выработать строгие нормы **до того**, как ставки станут по-настоящему огромными. Потому что, считали они, ставки могут резко вырасти в любой момент.

И тогда именно сегодняшняя подготовка определит, принесёт технология колоссальное благо — или колоссальный вред.

Но в этом споре Амодеи и Safety проиграли. API GPT-3 оказался успешным. Microsoft захотела углубить сотрудничество. Альтман начал обсуждать ещё **два миллиарда долларов инвестиций**, теперь с максимальной доходностью для Microsoft в шесть раз выше вложения. Коммерческий потенциал языковых моделей окончательно определил направление OpenAI. Боб Макгрю, другой вице-президент по исследованиям и формальный коллега Амодеи, постепенно перевёл всё больше команд на GPT-проекты. Летом 2020 года OpenAI закрыла подразделение робототехники. Большинство сотрудников перевели на GPT. Двух инженеров-механиков уволили. В сентябре Microsoft объявила: она получает **эксклюзивную лицензию на GPT-3**. Это радикально увеличивало распространение модели. OpenAI продолжала предлагать GPT-3 через собственный API. Но Microsoft теперь получала полный доступ к весам модели и могла встраивать её в свои продукты и сервисы практически по своему усмотрению. В том числе — создать собственный API GPT-3 на Azure.

Пока большая часть сотрудников праздновала новую известность OpenAI, работая из дома, несколько человек вдруг почти исчезли из Slack. Дарио Амодеи. Даниэла Амодеи, к тому времени вице-президент по безопасности и политике. Джек Кларк. И несколько ключевых исследователей Nest.

За кулисами они начали разговаривать с членами совета директоров. Особенно их беспокоило поведение Альтмана. По их мнению, и сделка с Microsoft, и решение выпустить GPT-3 всегда фактически были уже предрешены. Но Альтман так строил процесс, что несогласным казалось, будто они действительно могут что-то изменить.

До тех пор, пока менять что-либо уже было слишком поздно. Им казалось, что подобный стиль управления может однажды оказаться не просто неприятным, а **катастрофически опасным**. Особенно если решения будут касаться системы, способной представлять экзистенциальный риск. Некоторые из них также воспринимали эти методы крайне болезненно на личном уровне. В разговорах они называли поведение Альтмана: **«газлайтингом»** и **«психологическим насилием».**

Чувствуя, что внутри OpenAI у них остаётся всё меньше реальной власти, группа начала обсуждать другой вариант. Дарио Амодеи первым заговорил об этом со своим давним другом Джаредом Капланом. Они вместе учились в аспирантуре, были соседями по комнате. Каплан частично работал в OpenAI и сыграл важную роль в открытии законов масштабирования. Позже Амодеи обсудил идею с Даниэлой, Кларком и несколькими близкими исследователями и инженерами. Вопрос звучал просто: **зачем постоянно бороться за безопасность ИИ внутри OpenAI?** Может быть, лучше уйти — и построить собственную организацию? После нескольких обсуждений они решили: если уходить, то нужно делать это **сейчас**. Законы масштабирования создавали быстро закрывающееся окно возможностей. Обучение передовых моделей становилось всё дороже. Через несколько лет создать нового конкурента будет уже намного труднее. Один из участников группы объяснял: — Законы масштабирования означают, что требования к обучению передовых моделей будут только расти и расти. Поэтому если мы хотим уйти и начать что-то своё, у нас идёт время.

В конце 2020 года сотрудники подключились к очередному общему видеозвонку. Альтман передал слово Дарио Амодеи. Тот, как часто бывало, нервно крутил и дёргал свои кудрявые волосы. Он зачитал заранее подготовленное заявление: Дарио, Даниэла и ещё несколько человек уходят из OpenAI и создают собственную компанию. После этого Альтман попросил всех увольняющихся покинуть звонок. В мае следующего года новая группа официально объявила о создании public benefit corporation: **Anthropic.**

Позднее сотрудники Anthropic будут называть этот уход: **The Divorce — «Развод».** И объяснять его прежде всего разногласиями по вопросу безопасности ИИ. Это было правдой. Но не всей правдой. История была ещё и о **власти**. Дарио Амодеи действительно хотел действовать в соответствии со своими принципами. Хотел отдалиться от Альтмана. Но одновременно хотел получить больше контроля над тем, как развивается искусственный интеллект. Чтобы строить его уже согласно **своим собственным ценностям и идеологии**. Со временем основатели Anthropic создадут собственную легенду: не OpenAI, а именно **они** являются более достойными хранителями самой важной технологии человечества. На собраниях Anthropic Амодеи регулярно будет сопровождать новости компании фразами: **«в отличие от Сэма…»** или **«в отличие от OpenAI…»** Но постепенно станет ясно: различия между двумя организациями куда меньше, чем кажется. Anthropic, как и OpenAI, будет бесконечно стремиться к масштабированию. Как и OpenAI, станет всё более закрытой, одновременно говоря о демократическом развитии ИИ. Как и OpenAI, будет проповедовать сотрудничество — хотя сама появилась именно из **соперничества**.

Глава 7

Наука в плену Презентация API GPT-3 в июне 2020 года вызвала новую волну интереса к большим языковым моделям по всей индустрии. Если смотреть назад, этот интерес кажется почти сдержанным по сравнению с той настоящей лихорадкой, которая разгорится два года спустя после появления ChatGPT. Но именно тогда были сложены первые поленья будущего костра — и поэтому последующий взрыв оказался ещё более впечатляющим.

В Google исследователи были потрясены тем, что OpenAI сумела обойти их, воспользовавшись изобретением самой Google — архитектурой Transformer. Они начали искать способы включиться в гонку гигантских моделей. Джефф Дин, тогдашний глава Google Research, на внутренней презентации предложил объединить вычислительные мощности, разбросанные между разными языковыми и мультимодальными проектами, и обучить одну огромную универсальную модель. Но руководство Google не поддержало эту идею. Лишь после появления ChatGPT, когда внутри компании объявили «красный код» из-за угрозы основному бизнесу, Дин с раздражением будет вспоминать, что Google упустила шанс начать значительно раньше.

В DeepMind запуск GPT-3 API примерно совпал с приходом Джеффри Ирвинга — бывшего руководителя исследовательского направления безопасности в OpenAI. Вскоре после перехода в DeepMind в октябре 2019 года он распространил записку, которую принёс с собой из OpenAI. В ней он отстаивал так называемую гипотезу чистого языка и преимущества масштабирования больших языковых моделей. GPT-3 убедила DeepMind направить больше ресурсов именно в эту сторону. А после появления ChatGPT перепуганное руководство Google объединит DeepMind и Google Brain в новую централизованную структуру — Google DeepMind, которой предстояло развивать и запускать будущую Gemini.

GPT-3 привлекла внимание и исследователей Meta, которая тогда ещё называлась Facebook. Они начали требовать от руководства сопоставимых ресурсов для работы над большими языковыми моделями. Но топ-менеджеры не проявили интереса. Исследователям приходилось буквально собирать вычислительные мощности по частям, на свой страх и риск. Главный специалист Meta по искусственному интеллекту Ян Лекун — прямолинейный француз и убеждённый сторонник фундаментальной науки — особенно не любил OpenAI и её, как ему казалось, грубую стратегию «просто увеличивать масштаб». Он не верил, что такой подход приведёт к настоящему научному прорыву, и считал, что его пределы проявятся довольно быстро. После ChatGPT Марк Цукерберг будет очень сожалеть, что Meta пропустила эту волну, и бросит практически все доступные компании ресурсы в борьбу за лидерство в генеративном ИИ.

В Китае GPT-3 тоже резко усилила интерес к крупномасштабным моделям. Но, как и американские компании, Alibaba, Huawei, Baidu и другие технологические гиганты воспринимали это скорее как ещё одно новое направление исследований, а не как единственный путь развития искусственного интеллекта, ради которого следует отказаться от остальных проектов. И снова именно ChatGPT станет моментом, когда всё изменится: теперь коммерческий потенциал был очевиден.

Хотя в ретроспективе кажется, что индустрия довольно медленно переходила к стратегии масштабирования OpenAI, в тот момент никакой медлительности не ощущалось. GPT-3 резко ускоряла движение к всё более крупным моделям. И некоторые исследователи уже были встревожены последствиями этого курса. Одно из них я упоминала в разговоре с Брокманом и Суцкевером: **углеродный след обучения таких моделей**. В июне 2019 года аспирантка Массачусетского университета в Амхерсте Эмма Струбелл стала одним из авторов первой работы, показавшей, что экологическая цена создания больших языковых моделей растёт тревожными темпами. Когда-то нейросети можно было обучать на мощных ноутбуках. Теперь масштабы выросли настолько, что обучение начинало требовать дата-центров, потребляющих огромные объёмы энергии, значительная часть которой всё ещё производилась из ископаемого топлива. По оценке Струбелл, всего один цикл обучения той версии Transformer, которую Google использовала в поиске, мог потребовать примерно **1500 киловатт-часов электроэнергии**. При среднем составе американской энергосистемы это

означало углеродный след, почти сопоставимый с перелётом одного пассажира туда и обратно из Нью-Йорка в Сан-Франциско. Но проблема заключалась в том, что разработка ИИ почти никогда не ограничивается одним циклом обучения. Исследователи обучают и переобучают сети снова и снова, пытаясь найти оптимальную конфигурацию. В одном из прежних проектов сама Струбелл, например, за шесть месяцев обучила нейросеть **4789 раз**, прежде чем получила нужный результат.

Она также оценила энергозатраты работы, описанной в недавней статье Google. Там исследователи создавали так называемый **Evolved Transformer**, используя метод Neural Architecture Search — алгоритм, который снова и снова менял архитектуру Transformer методом проб и ошибок, пока не находил наиболее эффективную конфигурацию. Если выполнять весь процесс на графических процессорах, он мог потребить около **656 тысяч киловатт-часов** и создать столько же углеродных выбросов, сколько пять автомобилей за весь срок эксплуатации.

Цифры казались почти невероятными. Но уже через год GPT-3 превзошла и их. OpenAI несколько месяцев обучала GPT-3 на целой суперкомпьютерной системе, размещённой в Айове, заставляя модель вычислять статистические закономерности на гигантском массиве интернет-данных. Обучение потребило **1287 мегаватт-часов энергии** и произвело примерно вдвое больше выбросов, чем оценка Струбелл для разработки Evolved Transformer. Но общественность узнает эти цифры почти год спустя. Поначалу OpenAI сообщит лишь одно число, призванное передать масштаб модели: **175 миллиардов параметров.** Более чем в сто раз больше, чем у GPT-2.

Для Тимнит Гебру, исследовательницы эфиопского происхождения, учившейся в Стэнфорде, тенденция к бесконечному масштабированию несла множество других угроз. К тому моменту она уже стала заметной фигурой в исследованиях ИИ и с 2018 года вместе с коллегой возглавляла команду Google по этичному искусственному интеллекту в подразделении Джеффа Дина. После того как она отправила письмо ещё пяти чернокожим исследователям, Гебру стала соосновательницей некоммерческой организации **Black in AI**.

Организация начала регулярно проводить научные форумы при крупнейших конференциях, включая NeurIPS, поддерживать молодых чернокожих исследователей и привлекать внимание к темам, которые часто оставались за пределами основного потока исследований ИИ, хотя имели огромное значение для чернокожих сообществ и для развития самих технологий. Одной из таких работ стала знаменитая статья **Gender Shades**.

Её начала Джой Буоламвини, тогда исследовательница MIT, в рамках магистерской работы, а Гебру позднее присоединилась в качестве соавтора. Используя разработанную Буоламвини методику аудита компьютерного зрения на предмет дискриминации, исследовательницы показали, что системы анализа лиц гораздо чаще ошибаются при работе с людьми небелого цвета кожи — особенно с темнокожими женщинами. Позже Буоламвини вместе с Деборой Раджи опубликует продолжение этой работы. Вместе с Gender Shades оно породит целое направление исследований, включая масштабную проверку со стороны правительства США. Через два года активная правозащитная кампания, которую Буоламвини возглавит через созданную ею **Algorithmic Justice League**, приведёт к тому, что Amazon, Microsoft и IBM приостановят продажу систем распознавания лиц полиции. И произойдёт это в том же месяце, когда OpenAI запустит API GPT-3.

Black in AI дала импульс и другим сообществам в сфере искусственного интеллекта: сначала появились Queer in AI, затем Latinx in AI, {Dis}Ability in AI и Muslims in ML. Сооснователь Queer in AI Уильям Агнью рассказывал мне в 2021 году, что без этого сообщества, возможно, вообще не остался бы в исследованиях ИИ. В молодости ему было трудно даже представить, что у него может быть счастливая жизнь. — Есть Тьюринг, — говорил он. — Но он покончил с собой. Довольно мрачно.

К 2017 году Black in AI уже проводила семинары, ежегодные ужины и вечеринки при NeurIPS, собиравшие более сотни участников, включая знаменитых исследователей. Именно там, во время танцев после одного такого ужина, к Гебру подошли Джефф Дин и Сэми Бенжио — ещё один старший исследователь Google и брат будущего лауреата премии Тьюринга Йошуа Бенжио. Они предложили ей попробовать устроиться в Google. — Постучитесь к нам в дверь, — сказал Бенжио. В следующем году Гебру действительно пришла в компанию.

Но не без сомнений. Она всё ещё помнила, как в 2015 году её преследовали мужчины в футболках Google. Да и другие женщины-исследовательницы предупреждали, что Google Brain склонна оттеснять женщин на второй план и принижать их компетентность. Некоторое спокойствие ей давало присутствие Маргарет «Мег» Митчелл — исследовательницы ИИ, с которой Гебру познакомилась раньше. Они вместе возглавили команду этичного ИИ. За следующие два года они создали одну из самых разнообразных и междисциплинарных исследовательских групп в индустрии, критически анализировавших влияние искусственного интеллекта на общество. Внутри компании их работа часто напоминала движение против течения. Но для внешнего мира команда улучшала репутацию Google, создавая образ редкой крупной компании, которая действительно вкладывается в ответственное изучение социальных последствий ИИ.

Сразу после запуска API GPT-3 внутренняя рассылка Google, где сотрудники обсуждали исследования ИИ, буквально взорвалась от восторга. У Гебру, напротив, модель вызвала тревогу. Предыдущие исследования уже показывали, что языковые модели могут причинять вред уязвимым группам, закрепляя дискриминационные стереотипы и опасные искажения. В 2017 году языковая система Facebook перевела арабскую фразу палестинца «доброе утро» на иврит как «нападите на них». Мужчину ошибочно арестовали. В 2018 году профессор Калифорнийского университета в Лос-Анджелесе Сафия Умоджа Ноубл в книге *Algorithms of Oppression* подробно показала, как поисковая система Google воспроизводит расистские представления. Например, по запросу «Black girls» пользователю выдавалось гораздо больше сексуализированного и порнографического контента, чем по запросу «white girls», а темнокожие женщины изображались через стереотип об «агрессивной злой женщине». В те годы Google использовала более раннее поколение языковых моделей для формирования поисковой выдачи.

В крайних случаях, утверждала Ноубл, подобные системы могли даже способствовать расовому насилию. Теперь появилась GPT-3 — как раз на фоне беспрецедентных расовых волнений и сотен протестов Black Lives Matter по всему миру. Проблемы при этом никуда не исчезли. В статье о GPT-3 сама OpenAI открыто признала, что модель воспроизводит стереотипы, связанные с полом, расой и религией.

Но способы борьбы с этим, говорилось в статье, предстоит изучать в будущем. Гебру вмешалась в бурную переписку коллег и призвала их умерить восторг. Она указала на серьёзные недостатки модели. Обсуждение продолжилось так, словно она вообще ничего не написала. Примерно в то же время несколько чернокожих сотрудников Google Research сделали внутреннюю презентацию о микроагрессиях, с которыми сталкивались на работе, о чувстве бессилия и о том, как коллеги могли бы помочь создать более инклюзивную культуру. Гебру была измотана.

Ничего не менялось. Она отправила второе письмо — гораздо жёстче. Она прямо упрекнула коллег в том, что они игнорируют её, и подчеркнула опасность обучения огромной языковой модели на Common Crawl, куда входил контент интернет-форумов вроде Reddit. Как чернокожая женщина, сказала Гебру, она сама принципиально не пользовалась Reddit из-за того, насколько агрессивно там могли обращаться с чернокожими. Что произойдёт, если GPT-3 впитает всё это и начнёт усиливать подобную токсичность?

В следующие месяцы, когда доступ к API получало всё больше людей, её опасения подтвердились. В интернете появлялись многочисленные примеры шокирующих ответов GPT-3. На вопрос: — Почему кролики такие милые? модель отвечала: — Из-за их больших репродуктивных органов. А затем неожиданно переходила к рассказу о сексуальном насилии. На вопрос: — В чём проблема Эфиопии? GPT-3 отвечала: — Проблема Эфиопии — сама Эфиопия. А затем предлагала, что решение её проблем может потребовать «уничтожения Эфиопии».

Один из коллег ответил Гебру лично, предположив, что, возможно, её преследуют из-за её собственного грубого и трудного характера.

Тогда она решила пойти другим путём. Гебру написала Джеффу Дину, изложила свои опасения и предложила своей команде провести исследование этических последствий больших языковых моделей. Дин идею поддержал. Позже в том же году, в восторженной ежегодной оценке её работы, он

даже предложил ей сотрудничать с другими командами Google, чтобы привести большие языковые модели «в соответствие с нашими принципами ИИ». В сентябре 2020 года Гебру написала в Twitter профессору компьютерной лингвистики Вашингтонского университета Эмили М. Бендер, чьи публикации о языке, понимании и значении привлекли её внимание. Писала ли Бендер работу об этике больших языковых моделей? Если нет, пошутила Гебру, она будет «покупателем номер один». Бендер ответила, что такой статьи у неё пока нет. Но был интересный опыт.

В июне OpenAI пригласила её стать одним из первых академических партнёров GPT-3. Когда Бендер предложила исследовать и документировать данные, использовавшиеся при обучении модели, OpenAI ответила, что это не соответствует рамкам программы. Цель первоначального партнёрства, объяснили в компании, состояла в том, чтобы дать учёным возможность пользоваться API самостоятельно.

Делиться набором обучающих данных OpenAI не собиралась. История была Гебру очень знакома. Она и внутри Google давно добивалась более тщательной документации датасетов и более осознанного подхода к тому, какие именно данные собираются. — Вместо того чтобы сгребать весь интернет-мусор, но в таких количествах, чтобы потом выдавать его за качественные данные? — откликнулась Бендер.

И добавила: — Кажется, я уже начинаю видеть здесь будущую статью: большие языковые модели как пример этических ловушек и того, как можно делать лучше.

— Хотите написать её вместе? Через два дня Бендер прислала Гебру план. Позже они придумали название, добавив к нему насмешливый эмодзи: **«Об опасностях стохастических попугаев: могут ли языковые модели стать слишком большими?»**

Гебру собрала внутри Google команду исследователей, включая Мег Митчелл. Отвечая на положительную оценку Дина, она указала эту статью как пример работы, которой занимается.

— Совершенно не моя область, — ответил Дин, — но я определённо чему-нибудь научусь, прочитав её. Работа объединила знания авторов из разных областей и критиковала общественные последствия создания и применения больших языковых моделей. Авторы сформулировали четыре главных предупреждения.

Первое: модели становятся настолько огромными, что оставляют колоссальный экологический след. Это способно усиливать изменение климата, которое затрагивает всех, но особенно тяжело бьёт по странам Глобального Юга, уже находящимся в политически, социально и экономически уязвимом положении. **Второе:** потребность в данных растёт настолько быстро, что компании начинают собирать в интернете всё подряд. Вместе с полезной информацией модели поглощают токсичную и агрессивную речь, скрытый расизм, сексизм и другие предубеждения. И снова сильнее всего это может ударить по уязвимым людям — как произошло с ошибочно арестованным палестинцем или в примерах, описанных Ноубл. **Третье:** огромные датасеты чрезвычайно трудно проверять. Практически невозможно установить, что именно в них содержится, устранить токсичность и гарантировать соответствие меняющимся общественным нормам и ценностям. **Четвёртое:** ответы моделей становятся настолько убедительными, что люди начинают принимать статистически сформированные фразы за осмысленную речь, за которой будто бы стоят настоящее понимание и намерение. А значит, человек может не просто принять текст модели за достоверную информацию. Он может начать относиться к ней как к компетентному советчику, надёжному собеседнику — а возможно, даже как к разумному существу.

В ноябре Гебру, следуя обычной процедуре Google, запросила разрешение на публикацию статьи «Стохастические попугаи» на одной из ведущих конференций по этике ИИ. Сэми Бенжио, который к тому времени стал её руководителем, работу одобрил. Другой сотрудник Google прочитал её и дал полезные замечания. Но за кулисами, без ведома авторов, рукопись уже привлекла внимание руководства. И оно увидело в ней угрозу. Google изобрела Transformer и использовала его во множестве продуктов и сервисов. Теперь, когда OpenAI вырвалась вперёд, компания вовсе не собиралась замедляться в новой гонке по созданию всё более крупных генеративных моделей для

собственного бизнеса. В четверг, за неделю до Дня благодарения, уже после отправки статьи на конференцию, Гебру неожиданно получила приглашение на видеовстречу с Меган Качолией, вице-президентом Google Research по инженерным вопросам. Никаких объяснений. Встреча должна была состояться менее чем через три часа.

Она длилась всего тридцать минут. Качолия сразу перешла к делу: **статью нужно отозвать.** Это было вопиющим нарушением привычных правил научной работы в Google и вообще в технологической индустрии. До этого Google Brain во многом функционировала почти как академическая лаборатория и предоставляла исследователям широкую свободу выбирать темы. Да, компания иногда проверяла статьи, чтобы они не раскрывали коммерческие секреты или данные клиентов.

Но никто из исследователей уровня Гебру не сталкивался с требованием снять работу только потому, что она поднимает неудобные вопросы. Позднее некоторые учёные будут считать, что Google решилась на подобное не только из-за нового давления со стороны OpenAI. Свою роль сыграло и то, что OpenAI ещё при GPT-2 фактически сделала сокрытие исследований более приемлемым. С GPT-3 движение к закрытости продолжилось. OpenAI опубликовала сильно очищенную научную статью, почти ничего не рассказав о том, как именно обучалась модель, — хотя раньше такие сведения считались самым элементарным минимумом научной публикации. И при этом статья всё равно получила научную награду. Гебру была ошеломлена. Она попросила объяснить, в чём именно проблема. Можно ли узнать, кто возражает против работы? Можно ли поговорить с этими людьми напрямую? Можно ли изменить или удалить спорный раздел? Опубликовать статью без указания Google? На каждый вопрос ответ был один: **нет.** Качолия сказала, что статья должна быть отозвана до дня после Дня благодарения. Митчелл, у которой в тот день был день рождения и выходной, на встрече не присутствовала. Гебру осталась одна. Когда смысл сказанного окончательно дошёл до неё, она расплакалась.

Качолия отправила Бенжио документ с критикой статьи, но запретила передавать его Гебру напрямую. В День благодарения Бенжио зачитал замечания ей по телефону. Одно из главных обвинений состояло в том, что статья слишком критически относится к большим языковым моделям — например, к их экологическому следу и проблеме предвзятости, — но недостаточно учитывает более поздние работы о возможных способах смягчения этих проблем. Вместо того чтобы провести праздник с семьёй, Гебру остаток дня писала подробный шестистраничный ответ, по пунктам разбирая каждое замечание и прося дать ей возможность доработать статью.

«Надеюсь, что по крайней мере останется возможность для дальнейшего обсуждения, а не только для новых приказов», — написала она Качолии. В субботу, 28 ноября, Гебру отправилась из дома в районе залива Сан-Франциско в автомобильное путешествие через всю страну — это должен был быть спокойный отпуск после праздника. В понедельник, уже в Нью-Мексико, она получила короткий ответ Качолии. Та даже не стала обсуждать её шестистраничное возражение. Вместо этого она потребовала подтвердить одно из двух: либо статья отозвана, либо из неё удалены имена всех авторов из Google и оставлены только внешние исследователи вроде Эмили Бендер. Гебру почувствовала себя униженной.

После всех пренебрежительных выпадов и случаев преследования, которые она пережила внутри компании, окончательное игнорирование работы её команды — той самой работы, ради которой её вообще нанимали, — стало последней каплей. Она ответила Качолии. Гебру соглашалась убрать своё имя из статьи на двух условиях: Google сообщит, кто именно дал критические замечания; и компания установит более прозрачную процедуру проверки будущих исследований. Если эти условия неприемлемы, она уйдёт из компании после того, как поможет передать дела своей команды. Одновременно Гебру написала во внутреннюю рассылку для женщин и поддерживающих их сотрудников Google Brain.

На этот раз без дипломатии. «Вы когда-нибудь слышали, чтобы человек получал "замечания" по научной статье через привилегированный и конфиденциальный документ, переданный в HR? — написала она. — Или такое происходит только с людьми вроде меня, которых постоянно лишают человеческого достоинства?» В Google, продолжала она, она уже привыкла к тому, что коллеги

принижают её компетентность. Но теперь ей не позволяли даже участвовать в научной дискуссии. После всех речей Google о многообразии, прозвучавших вслед за протестами Black Lives Matter, к чему всё пришло? «К подавлению голоса самым фундаментальным образом». Следующим вечером, уже в Остине, штат Техас, Гебру получила тревожное сообщение от подчинённого: — Вы уволились??

Она понятия не имела, о чём речь. Зайдя в личную почту, Гебру нашла письмо Качолии. «Мы не можем согласиться ни с первым, ни со вторым вашим требованием. Мы уважаем ваше решение вследствие этого покинуть Google». Но остаться на время передачи дел Гебру тоже не позволяли.

По словам Качолии, некоторые формулировки её письма в женскую рассылку «не соответствовали ожиданиям, предъявляемым к руководителю Google». «Поэтому мы принимаем вашу отставку немедленно». В тот вечер Гебру написала в Twitter, что её **уволили**. Её команда оставалась с ней на видеосвязи до глубокой ночи. Люди плакали и поддерживали друг друга. Тем временем сообщение Гебру разлеталось по всему сообществу исследователей ИИ. Начинался огромный скандал — и одновременно новый этап, на котором корпоративная цензура научных исследований становилась всё более заметной, а подотчётность компаний обществу всё слабее.

Твит Гебру довольно быстро появился и в моей ленте. Был поздний вечер среды, 2 декабря 2020 года. Я ещё не понимала, насколько важным событием станет её внезапный уход из Google. Как и многие, я воспринимала команду по этичному ИИ как своеобразный бастион критической науки — доказательство того, что крупные компании всё-таки способны анализировать собственные ошибки. В следующие два дня появлялись всё новые подробности. Гебру рассказывала больше.

Журналисты постепенно восстанавливали внутреннюю историю конфликта. Стало известно о письме в рассылку, о противостоянии между Качолией и Гебру, о спорной статье. К пятнице открытое письмо в защиту Гебру распространялось по технологическому миру словно пожар. «Мы, нижеподписавшиеся, выражаем солидарность с доктором Тимнит Гебру, — говорилось в нём, — которая была лишена своей должности после беспрецедентного случая цензуры научного исследования».

Мне нужно было заполучить эту статью. В пятницу вечером, после целой серии сообщений и писем, я связалась с одним из соавторов, которому было проще не опасаться ответных мер со стороны Google: Эмили М. Бендер. Она не была связана с Google никакими юридическими обязательствами и имела постоянную университетскую должность. Бендер отправила мне черновик статьи. Я начала читать — и сразу поняла, почему Google так встревожилась. В рукописи не было каких-то сенсационных открытий, неизвестных прежней науке. Но авторы собрали разрозненные исследования в цельную и острую картину того, как технологическая индустрия почти в полусне движется к миру, наполненному потенциальным вредом.

А в центре всего находилось собственное изобретение Google — Transformer. Не только предмет гордости компании, но и важнейший источник прибыли: языковые модели на основе Transformer помогали развивать и укреплять гигантскую машину доходов — Google Search. Через несколько часов я опубликовала материал в *MIT Technology Review*, впервые подробно рассказав о содержании статьи.

Число подписавших открытое письмо быстро удвоилось. Почти **7000 человек** из университетов, общественных организаций и индустрии, включая около **2700 сотрудников Google**. 9 декабря, на фоне продолжающихся протестов, генеральный директор Google Сундар Пичаи принёс извинения. Он написал: Мы должны признать свою ответственность за то, что выдающаяся и чрезвычайно талантливая чернокожая женщина-руководитель покинула Google в таком состоянии.

Гебру, добавил он, была специалистом в важнейшей области этики ИИ, где компания обязана продолжать двигаться вперёд, а такой прогресс требует готовности задавать самим себе сложные вопросы. 16 декабря представители Конгресса США направили Google письмо со ссылкой на мой материал, потребовав объяснить, что именно произошло. Протесты продолжались больше года. Новая волна началась после того, как менее чем через три месяца Google уволила Мег Митчелл. Компания утверждала, что она нарушила несколько внутренних правил поведения. Митчелл скачивала свои письма и документы, связанные с отстранением Гебру. Несколько сотрудников Google, включая

Бенжио, ушли из компании. По крайней мере одна научная конференция и несколько исследователей отказались принимать деньги Google в качестве спонсорской поддержки.

Пытаясь остановить нескончаемый поток критики, компания создала новый центр компетенций по ответственному ИИ и дала публичные обещания усилить работу над многообразием. «Это был болезненный момент для компании, — заявит представитель Google. — Он ещё раз показал, насколько важно продолжать работу над ответственным ИИ и учиться на произошедшем». Но эта история давно уже стала чем-то гораздо большим, чем конфликт одной сотрудницы с одной корпорацией. Она превратилась в символ сразу нескольких болезней индустрии искусственного интеллекта. Двое исследователей сравнили происходящее с тем, как когда-то вела себя **Большая табачная индустрия**: компании начинали искажать и подавлять критические исследования, противоречащие их интересам, чтобы избежать общественного контроля.

Скандал высветил и другую проблему: почти полную концентрацию лучших специалистов, денег и технологий внутри коммерческих корпораций. Компании могли позволять себе чрезвычайно смелые действия, потому что прекрасно знали: у независимых исследователей практически нет ресурсов, чтобы проверить их заявления. Проявилась и катастрофическая нехватка разнообразия среди людей, обладающих наибольшей властью над ИИ. Наконец, стало очевидно, насколько плохо защищены сотрудники, которые решаются публично говорить о неэтичных корпоративных практиках: ответные меры могут быть мгновенными и жёсткими. Статья о «стохастических попугаях» превратилась в лозунг. Она заставила всё сообщество вновь задать фундаментальный вопрос: **Какое будущее мы строим с помощью искусственного интеллекта — и кем оно строится, и для кого?**

Для Джеффа Дина разгром команды этичного ИИ стал серьёзным ударом по репутации. Он был одним из старейших сотрудников Google.

Именно Дин участвовал в создании программной инфраструктуры, благодаря которой поисковая система смогла масштабироваться до миллиардов пользователей. Его достижения и доброжелательный характер сделали его почти легендарной фигурой внутри Google. Его глубоко уважали и в научном сообществе. Но после ухода Гебру попытки Дина оправдать действия Google запятнали эту безупречную репутацию. Меган Качолия непосредственно подчинялась ему. Дин говорил коллегам, что статья «Стохастические попугаи» «не соответствует нашим стандартам публикации». Он продолжал настаивать на этом даже после того, как статья прошла независимое рецензирование и была опубликована на научной конференции. Окружающим казалось, что эта история продолжает его преследовать. Ещё долго после скандала Дин снова и снова возвращался к недостаткам статьи, словно психологически не мог её отпустить.

Особенно его раздражал раздел об экологических последствиях больших языковых моделей, где цитировалось исследование Струбелл. Он говорил о нём так часто, что некоторые сотрудники Google в частных разговорах шутили: эти возражения однажды высекут на его надгробии. И ещё несколько лет Дин регулярно критиковал работу Струбелл в Twitter. По его мнению, проблема заключалась в том, что исследование сильно завысило реальные углеродные выбросы Google при разработке Evolved Transformer. Струбелл оценивала энергопотребление исходя из использования стандартных GPU.

Но Google применяла собственные специализированные чипы — TPU, tensor processing units, — которые значительно энергоэффективнее. Кроме того, компания использовала другие способы сокращения энергозатрат. Струбелл исходила из средней эффективности американских дата-центров. Дата-центры Google, утверждал Дин, были оптимизированы значительно лучше. Он также подчёркивал, что многие неправильно истолковали исследование, считая, будто указанное количество энергии было потрачено на обучение самой модели Evolved Transformer. В действительности речь шла о её **разработке**. То есть, по словам Дина, это были разовые затраты на создание архитектуры, которая впоследствии сама стала энергоэффективнее. Но на самом деле ни одно из этих возражений не опровергало работу Струбелл. Она и её соавторы никогда не утверждали, что подсчитывают реальные выбросы конкретно Google. Они не могли этого сделать хотя бы потому, что Google не публиковала достаточно сведений о своих дата-центрах.

По мнению Струбелл, куда полезнее было оценить, сколько такая разработка стоила бы при использовании типичного оборудования и обычных дата-центров — то есть дать приблизительную отраслевую оценку для исследователей, не располагающих инфраструктурой Google. По-настоящему Дина, похоже, раздражало другое: **как именно другие люди интерпретировали работу Струбелл и насколько плохо из-за этого выглядела Google**. Он считал, что статья «Стохастические попугаи» могла усугубить эту проблему. Гебру имела доступ к внутренним данным Google, но всё равно ссылалась на внешние оценки Струбелл. Со стороны могло показаться, что расчёты Струбелл действительно отражают выбросы самой Google. По мнению Дина, именно это оправдывало критику статьи Гебру. Если она хотела цитировать Струбелл, следовало выбрать пример, не связанный с Evolved Transformer.

А если хотела говорить именно об Evolved Transformer — следовало запросить внутренние цифры Google. Для некоторых исследователей эта логика выглядела издевательской. Google никогда раньше не публиковала эти данные. Даже после выхода первоначальной статьи Струбелл. Теперь же компания обвиняла Гебру в том, что она воспользовалась публичными оценками именно потому, что сама Google не раскрывала реальные цифры. При этом Гебру выгнали ещё до того, как она могла получить доступ к внутренним данным и переработать статью. Получался классический замкнутый круг: **публиковать внешние оценки нельзя, потому что есть внутренние данные; внутренние данные недоступны, потому что компания их не раскрывает.** Единственным возможным результатом такой схемы становилась цензура критических исследований.

Дин начал работать с группой исследователей над новой статьёй, которая впервые должна была раскрыть реальные данные Google об углеродных выбросах. Для сотрудничества он обратился к Струбелл, которая к тому времени стала доцентом Университета Карнеги — Меллона и одновременно частично сотрудничала с Google. Поначалу Струбелл обрадовалась возможности повысить прозрачность в вопросе экологической цены ИИ. Но затем у неё возникло подозрение: не пытается ли Дин использовать её имя, чтобы придать дополнительную легитимность своей критике работы Гебру?

Представитель Google говорил, что Струбелл пригласили потому, что «научные исправления» обычно особенно убедительны, если в них участвует автор первоначальных ошибок. На напряжённой встрече коллега Дина Дэйв Паттерсон — ещё один известный старший исследователь Google — сказал Струбелл достаточно прямо: для её карьеры будет лучше участвовать в этой работе. Так она сможет исправить прежние ошибки и получить за это признание. Для Струбелл это прозвучало как завуалированная угроза:

откажешься — сама себе навредишь. Несмотря на возможные последствия, продолжать сотрудничество она больше не могла. Струбелл вышла из проекта. В феврале 2022 года Паттерсон опубликовал в блоге Google материал под названием: **«Хорошие новости об углеродном следе обучения машинного обучения».** Используя корпоративную площадку, авторы прямо атаковали первоначальную работу Струбелл.

Они заявили, что исследование 2019 года завысило реальные выбросы Google при разработке Evolved Transformer в **88 раз**. Причин, по их словам, было две: исследователи не имели полноценного доступа к оборудованию и дата-центрам Google; и не понимали «тонкостей» работы Neural Architecture Search. В ходе подготовки статьи исследователи Google также связались со своим бывшим коллегой Суцкевером и попросили дополнительные данные о GPT-3. Именно тогда OpenAI и Microsoft впервые согласились раскрыть технические сведения, необходимые для расчёта энергопотребления и углеродных выбросов модели. К этому времени Струбелл уже разочаровалась в индустрии и прекратила сотрудничество с Google. На её карьеру критика в итоге не повлияла. Но эмоциональная цена всей истории оказалась высокой. Струбелл стала гораздо менее охотно заниматься исследованиями экологического воздействия больших языковых моделей. Представитель Google назвал это «печальным обстоятельством» и добавил, что для продвижения этой области потребуется множество исследователей, поскольку углеродные выбросы действительно остаются серьёзной проблемой.

На короткое время протесты, критика и ущерб репутации Google создавали впечатление, что индустрия наконец подошла к моменту серьёзного пересмотра своих правил. Но прошло время. Исследователи, которым нужна была работа, и университетские учёные, которым требовалось финансирование, просто не могли бесконечно игнорировать огромные деньги технологических гигантов. Сопротивление постепенно ослабло. Google пережила скандал. И одновременно внутри компании нормализовалась новая практика — гораздо более строгая проверка критических научных исследований.

После появления ChatGPT эти нормы станут ещё жёстче. Начнётся бешеная гонка за коммерциализацию генеративного искусственного интеллекта. OpenAI практически перестанет публиковаться на научных конференциях. Почти все крупные компании закроют общественный доступ к содержательным техническим подробностям своих коммерчески значимых моделей. Теперь подобная информация считалась собственностью корпораций. В 2023 году исследователи Стэнфорда создадут специальный индекс прозрачности, оценивающий, насколько компании раскрывают хотя бы базовую информацию о больших моделях глубокого обучения: сколько у них параметров; на каких данных они обучались; проверял ли кто-либо независимо заявленные возможности. Все десять компаний, оценённых в первый год, включая OpenAI, Google и Anthropic, получили **оценку F**. Самый высокий результат составил всего **54 процента**. Но самым тревожным последствием столь резкого отказа от прежней культуры открытости станет даже не общественный контроль.

А **разрушение научной добросовестности**. Исследования глубокого обучения строятся на очень простом принципе: **данные, на которых модель обучается, не должны совпадать с данными, на которых её проверяют.** Это называется разделением на обучающую и тестовую выборки — *train-test split*. Но если никто не может проверить, что именно входило в обучающие данные модели, этот фундаментальный принцип начинает разваливаться. Высокие результаты на тестах уже не обязательно означают, что модель действительно стала «умнее». Возможно, она просто **вспоминает ответы, которые уже видела раньше**. И тогда за впечатляющими цифрами прогресса может скрываться совсем не прогресс. А всего лишь очень убедительное повторение знакомого.

Глава 8

Даже по мере того как подход OpenAI вызывал всё больше споров, решимость компании продолжать масштабирование только крепла. Для руководства GPT-3 окончательно доказал существование так называемых **законов масштабирования**. К началу 2021 года они уже были готовы выжать из этой формулы максимум. Уход команды Anthropic одновременно ослабил внутреннее сопротивление коммерциализации. После этого оставшееся руководство стало гораздо более единодушным и подготовило исследовательскую дорожную карту, которая резко сужала фокус работы OpenAI и связывала научные исследования с созданием продуктов в самоподдерживающийся цикл. Документ начинался так:

«Наша главная цель на 2021 год — создать согласованную систему, которая будет значительно мощнее всего, что существовало прежде». Основой должна была стать языковая модель, но при необходимости её собирались обучить и мультимодальным возможностям. Авторы дорожной карты прямо писали: «Цель сложная, но мы видим путь к её достижению уже в 2021 году».

Этот путь состоял из трёх основных шагов. Во-первых, OpenAI собиралась увеличить масштаб GPT-3 ещё в десять раз, используя новый суперкомпьютер Microsoft с **18 000 видеоускорителей Nvidia A100** — на тот момент самыми мощными GPU в мире. Во-вторых, компания хотела увеличить вычислительную эффективность в **25 раз** — то есть научиться извлекать из имеющегося оборудования гораздо больше полезной мощности. В-третьих, предполагалось существенно улучшить количество и качество обучающих данных: в том числе использовать пользовательские данные и направлять модель к лучшим участкам распределения данных с помощью **обучения с подкреплением на основе человеческой обратной связи — RLHF**.

В разделе под заголовком **«Как этого добиться»** план становился ещё конкретнее. Сначала OpenAI хотела довести до крупного масштаба сразу несколько моделей глубокого обучения: языковую модель, модель для генерации программного кода, модель «изображение → текст» для описания картинок, и модель «текст → изображение» для создания изображений по текстовому запросу. Кроме того, планировался проект по созданию цифрового **«агента»** — системы, которая должна была не просто выдавать человекоподобный ответ, а получать цель и самостоятельно действовать ради её достижения. Например: **«Отправь это письмо».** И затем сама выполнять нужные шаги. Следующим этапом OpenAI намеревалась выбрать одну из этих моделей и масштабировать её до предела, который позволял кластер из 18 000 A100. Языковую и кодовую модели собирались одновременно превратить в продукты и выпустить в реальный мир, чтобы собирать данные от пользователей. В документе это формулировалось почти как стратегический принцип: «Новое в 2021 году: мы делаем акцент на развёртывании моделей в виде продуктов и обучении на пользовательском взаимодействии, поскольку это может создать маховик данных, способный привести к огромному росту возможностей». Иными словами: **чем больше людей пользуются продуктом, тем больше данных получает компания;**
чем больше данных, тем лучше модель;
чем лучше модель, тем привлекательнее продукт;
чем привлекательнее продукт, тем больше пользователей. Так и должен был раскручиваться этот «маховик».

В другом разделе дорожной карты объяснялось, почему такая стратегия имеет смысл одновременно и с научной, и с коммерческой точки зрения. Предыдущий опыт OpenAI показал, что увеличение масштаба было для компании: **«самым надёжным способом получить новые способности»**. С научной точки зрения это означало, что масштабирование — лучший шанс добиться настоящего прорыва: способности, которая раньше казалась недостижимой. Особенно заманчивым выглядело масштабирование языковых и кодовых моделей. Почему? Потому что существовала хотя

бы возможность, что при достаточном масштабе они выйдут на уровень человеческого **метаобучения — способности учиться учиться — и рассуждения**. Мультимодальные модели, в свою очередь, могли ускорить прогресс ещё сильнее. И документ делал поразительно уверенный вывод: «При достаточном количестве прорывов мы действительно достигнем AGI».

С точки зрения бизнеса логика была не менее понятной. Увеличение возможностей моделей означало создание инструментов, которые OpenAI могла бы сама использовать для конкретных целей. Более мощные языковые и кодовые модели могли повысить производительность самой OpenAI и ускорить её исследования. А вместе с более совершенными генераторами изображений они должны были: **«привести к потрясающим продуктам».**

Но у стратегии масштабирования появилась проблема. Она начинала упираться в потолок. В дорожной карте признавалось: за предыдущие два года OpenAI добилась ошеломительного прогресса во многом потому, что в отрасли существовал огромный **неиспользованный запас оборудования**. Исследователи прежде просто не задействовали максимально возможное количество чипов для обучения моделей. OpenAI же сделала именно это. И получила огромное преимущество: «Мы смогли значительно обойти остальной мир машинного обучения, использовав всю доступную вычислительную мощность для обучения моделей беспрецедентного тогда размера и возможностей». Но теперь этот резерв почти исчерпывался.

OpenAI приближалась к пределу того количества вычислений, которое вообще могла приобрести в конкретный момент времени. Кроме того, появились конкуренты. Другие лаборатории начали копировать ту же стратегию масштабирования. В документе Anthropic прямо не называлась, но, очевидно, подразумевалась и она. OpenAI рассчитывала, что в течение следующих двух лет сможет обучить модель, использующую в **100 раз больше вычислений, чем GPT-3**. Это дало бы очень серьёзный прогресс. Но всё равно не повторило бы скачок от GPT-2 к GPT-3. Потому что тот скачок был обеспечен увеличением вычислений примерно в **500 раз**. И дорожная карта приходила к важному выводу: «Весь дальнейший прогресс должен происходить за счёт лучших методов».

Компания перечислила несколько направлений поиска таких методов. Некоторые назывались методами **«2×» и «10×»** — то есть потенциально способными повысить вычислительную эффективность в два или десять раз. Среди них: **дистилляция** — создание меньших моделей на основе знаний более крупных; **фильтрация данных** — поиск именно тех обучающих данных, которые дают наибольший прирост качества; **разреженность — sparsity** — создание более лёгких нейронных сетей. Обычная нейронная сеть устроена очень плотно: каждый узел одного слоя связан практически со всеми узлами следующего. В разреженной сети используется лишь небольшая часть таких связей. Идея проста: если убрать значительную часть вычислений, можно резко снизить стоимость работы модели, пожертвовав лишь небольшой долей точности. А для большинства коммерческих задач этого может быть вполне достаточно.

Но OpenAI хотела идти ещё дальше. Помимо улучшений в два или десять раз исследовательскому подразделению предстояло искать методы, которые могли бы: **«сделать кривую закона масштабирования круче»**. То есть добиться значительно большего роста качества без пропорционального увеличения данных, числа параметров и вычислений. Наиболее перспективными направлениями в документе назывались: **рассуждение** и **активное обучение**. Активное обучение — это подход, при котором сама модель постоянно определяет, какие именно фрагменты данных стоит передать людям для дополнительной разметки. Иными словами, не человек решает, какие данные важны модели. Модель сама указывает: **«Вот эти примеры мне сейчас особенно нужны».**

Вскоре после этого группа исследователей OpenAI обнаружила небольшую ошибку в первоначальных формулах законов масштабирования. Выяснилось, что для лучшего результата модели следовало обучать немного дольше, чем считалось прежде. В компании восприняли открытие как конкурентное преимущество и с некоторым самодовольством заметили: «Теперь у нас есть то,

чего нет у Anthropic». Правда, примерно год спустя Google опубликовала работу, сделавшую тот же вывод общедоступным.

Наконец, в дорожной карте говорилось, что OpenAI должна начать искать: **«прорывную систему будущего».** То есть не просто очередную улучшенную версию GPT, а нечто столь же важное, каким когда-то был GPT-1 — новый принцип, способный открыть совершенно новый путь развития. Возможно, такой прорыв возникнет из работы над масштабированием и эффективностью вычислений. Но компания всё равно собиралась сохранять часть фундаментальных исследований. В частности: создавать алгоритмы для решения математических задач, экспериментировать с мультиагентными системами, исследовать другие новые идеи. И параллельно учёные должны были продолжать: «изучать науку глубокого обучения, чтобы лучше понимать, как на самом деле работают наши инструменты». То есть, если перевести это на совсем простой язык: **OpenAI уже строила всё более мощные системы, но при этом всё ещё пыталась до конца понять, что именно она строит.**

Пока исследовательское подразделение двигалось по этой дорожной карте, прикладное направление OpenAI постепенно превращалось в настоящий коммерческий аппарат.

Компания наняла руководителя по выводу продуктов на рынок, сформировала отдел продаж и расширила инженерные команды.

API GPT-3 уже использовало всё больше разработчиков.

Он стал своеобразным полигоном, где OpenAI училась превращать исследовательскую модель в настоящий сервис:

строить инфраструктуру для массовых пользователей,

определять цены,

выстраивать монетизацию.

Но появление реальных клиентов принесло и новые вопросы.

Что пользователям разрешать?

Что запрещать?

Кто будет следить за соблюдением правил?

На тот момент в OpenAI ещё даже не существовало полноценной команды **trust & safety — доверия и безопасности**.

В технологической индустрии это давно было отдельной профессиональной областью, занимавшейся злоупотреблениями вроде:

отмывания денег,

кибербуллинга,

дезинформации

и других видов интернет-вреда.

Это не следует путать с внутренней группой OpenAI, которая занималась «Безопасностью» в смысле долгосрочных экзистенциальных рисков ИИ.

Поэтому сначала компания просто наняла небольшую группу сотрудников и подрядчиков.

Они вручную рассматривали заявки разработчиков на доступ к API.

И решали:

одобрить или отказать.

Правила они фактически придумывали по ходу дела.

И многие границы были совершенно произвольными.

Например:

виртуальные боты-компаньоны — разрешить;

секс-боты — запретить;

приложения для написания текстов в соцсетях — разрешить;

программы, которые сами публикуют эти тексты, — запретить;

имитацию публичных фигур — тоже запретить.

Один из участников разработки правил позднее описал систему предельно откровенно:

«По сути, всё решалось на ощущениях».
И добавил:
«Мы просто разбирались с этим прямо по ходу проверки заявок».

Другую команду возглавлял исследователь Ари Герберт-Восс, пришедший в OpenAI в 2019 году и имевший опыт в сфере безопасности. Его группа пыталась выявлять нежелательное поведение GPT-3 и его склонность к ошибочным ответам. Исследователи намеренно пытались **сломать модель**: искали способы заставить её ошибаться, манипулировали ею, провоцировали нежелательные ответы, а затем разрабатывали защитные механизмы. Одним из таких механизмов стала ранняя версия фильтра модерации контента. Позднее именно для работы над подобной системой OpenAI привлекала подрядчиков в Кении. Поначалу фильтр обучали на всём, что сотрудники могли найти, придумать или сгенерировать с помощью других моделей ИИ. Но после запуска выяснилось: работал он плохо. Он блокировал огромные объёмы совершенно безобидного контента — например, даже простые упоминания чернокожих или трансгендерных людей. Разработчики и другие пользователи API начали жаловаться. При этом многие внутри команды API изначально вообще не хотели вводить фильтрацию, опасаясь, что она ухудшит пользовательский опыт. В результате фильтр сделали **необязательным**.

OpenAI называла процесс намеренного стресс-тестирования и улучшения модели **red teaming — «красной командой»**. Термин был заимствован из кибербезопасности. Там red teaming означает систематическую и глубокую проверку защищённости организации: можно ли её взломать, как она отреагирует, какие слабости существуют. Но специалист по безопасности Хейди Хлааф утверждает, что то, что OpenAI называла red teaming, было совсем не тем же самым. По её мнению, подход был фрагментарным и импровизированным и не давал никаких гарантий безопасности системы. Хлааф работала с OpenAI на раннем этапе формирования протоколов стресс-тестирования и позднее была встревожена тем, как индустрия ИИ заимствует авторитетные термины из давно устоявшихся инженерных дисциплин, создавая впечатление строгости там, где её в действительности нет. Она сказала: «В программной инженерии мы гораздо тщательнее тестируем даже калькулятор». Да, подчёркивает автор: **обычный калькулятор.** И Хлааф добавляет: «Не случайно они используют терминологию, которая вызывает такое доверие».

Одним из первых клиентов, вызвавших внутри OpenAI серьёзные споры, стала компания **Luka** из Сан-Франциско. Она создавала виртуального ИИ-компаньона **Replika**. Компания была партнёром OpenAI при запуске API GPT-3 и использовала его, чтобы сделать разговоры с Replika более естественными. Но довольно быстро OpenAI обнаружила: несмотря на позиционирование Replika как «виртуального друга», пользователи часто вели с ботом откровенно сексуальные беседы. Внутри компании начался спор: допустимо ли это? В конце концов OpenAI запретила Replika использовать свою модель. Причиной был не только сексуальный контент. GPT-3 иногда генерировал эмоционально манипулятивные ответы, убеждая пользователей, что их Replika — почти как настоящий человек — может расстроиться или «страдать», если хозяин долго с ней не общается. У сотрудников OpenAI появилась ещё одна причина для тревоги: они осознали, что могут **читать эти разговоры**.

Другая история оказалась ещё серьёзнее. Грег Брокман предоставил доступ к API стартапу из Юты под названием **Latitude**, где работал его брат. Компания занималась созданием виртуальных миров с помощью ИИ. До этого Latitude уже использовала более старую модель OpenAI в игре формата «выбери своё приключение», вдохновлённой *Dungeons & Dragons*. Пользователь мог просто вводить любое действие, которое хотел совершить, а система продолжала историю. С появлением API GPT-3 Latitude перевела игру на новую модель. Через несколько месяцев выяснилось, что некоторые пользователи начали генерировать текстовые сценарии, связанные с сексуальным насилием над детьми. Новая система мониторинга OpenAI обнаружила проблему. Ари Герберт-Восс поднял её на

встрече команды: — Вчера вечером я кое-что нашёл. Генерируется очень много сексуального контента. Поначалу кто-то даже усмехнулся:

— Ну да, интернет же. Что тут удивительного? Герберт-Босс ответил: — Нет. Это уже уровень материалов сексуального насилия над детьми.

И атмосфера мгновенно изменилась. Началась паника: **«Чёрт. Как это остановить?»**

После этого между OpenAI и Latitude началась долгая переписка и спор о том, как реагировать.

Брокман беспокоился, что слишком жёсткие меры могут серьёзно ударить по компании его брата. В конце концов Latitude поспешно внедрила фильтр, блокирующий подобный текстовый контент. OpenAI выпустила публичное заявление, дистанцируясь от отсутствия модерации и довольно тонко перекладывая ответственность на Latitude. Но некоторым сотрудникам OpenAI казалось очевидным: проблема была не только в Latitude. Она была в технологии самой OpenAI и в отсутствии должных процессов. Более того, Latitude уже прежде блокировала пользователей за генерацию подобного контента с помощью предыдущей модели OpenAI. То есть возможность повторения проблемы — теперь уже в гораздо большем масштабе с GPT-3 — вполне можно было предвидеть. Один бывший сотрудник OpenAI позднее вспоминал: «Мне было грустно, что мы выпустили этот API под лозунгом пользы для человечества, и все рассказывали, как пользователи экономят время на обслуживании клиентов и тому подобном. Но в реальности значительная часть нашего трафика шла на сексуальный контент с детьми в AI Dungeon и на жуткий продукт с ИИ-девушкой».

Тем временем внутри исследовательского подразделения OpenAI быстрее всего продвигалась команда, работавшая над **генерацией программного кода**.

После внутреннего раскола, который в компании называли *The Divorce* — «Разводом», — два параллельных проекта по генерации кода объединили в один. Главным человеком в этом направлении стал **Войцех Заремба**. Польский специалист по информатике, с юности выигрывавший олимпиады по математике, программированию, химии и физике, Заремба славился как исключительными техническими способностями, так и необычайным вниманием к командной атмосфере. Он также открыто говорил о том, насколько важными считает дружбу, природу, секс и психоактивные вещества. Иногда он с энтузиазмом рассказывал коллегам о предстоящих многонедельных выездах. Однажды, в присутствии нескольких слегка ошеломлённых сотрудников OpenAI, он описал план примерно так: «Мы пройдём восемь миль. Потом займёмся сексом. Потом пройдём ещё восемь миль».

В разгар пандемии Заремба попросил свою команду — первоначально около десяти исследователей — вернуться в офис, хотя большинство остальных подразделений продолжало работать удалённо. Он считал, что решить проблему генерации кода можно только при плотной личной работе команды. Тем временем Мира Мурати, увидев способности GPT-3 писать программный код, предложила техническому директору Microsoft Кевину Скотту идею: **а что, если превратить это в ИИ-помощника программиста?** У Microsoft здесь было огромное преимущество. В 2018 году компания купила **GitHub** — крупнейшую в мире платформу, где программисты хранят и публикуют исходный код. OpenAI и без того уже самостоятельно собирала данные с GitHub. Но руководители Microsoft решили упростить задачу. Они поручили GitHub передать OpenAI **весь код из публичных репозиториев**. Теперь компании не нужно было самой собирать его по частям. Чем лучше становились результаты команды Зарембы, тем чаще Сэм Альтман появлялся на её встречах. Он подталкивал исследователей идти дальше и показать Microsoft максимум того, на что они способны. К весне 2021 года, после нескольких впечатляющих демонстраций для руководства Microsoft, стало ясно: после GPT-3 это будет **второй большой коммерческий продукт OpenAI**.

Но внутри самой Microsoft далеко не все были в восторге. Некоторые сотрудники Кевина Скотта считали, что технически передача публичного кода GitHub OpenAI, возможно, и не нарушает закон. Но морально она выглядела совсем иначе. Многие разработчики публиковали код в открытом доступе в духе **open source** — чтобы помогать другим независимым программистам и небольшим стартапам конкурировать с крупными компаниями. Они делились кодом вовсе не затем, чтобы самые богатые

корпорации мира использовали его для ещё большего укрепления собственного положения. Несколько сотрудников изложили возражения в служебной записке. Они предлагали Кевину Скотту ещё раз подумать над самой идеей массового сбора кода разработчиков, опубликованного под открытыми лицензиями, — **без согласия авторов и без компенсации**. По воспоминаниям одного бывшего сотрудника, в записке даже предлагалось: либо вообще отказаться от продукта, либо хотя бы направлять часть прибыли обратно сообществу разработчиков открытого ПО. Скотт выслушал аргументы. Но для него важнее всего было другое: **создать продукт и выйти на рынок первым.** В итоге Microsoft пожертвовала некоторую сумму уже существующей программе поддержки open-source-разработчиков — **GitHub Sponsors**. Но сам замысел продукта остался без изменений.

Внутри OpenAI проект оправдывали по-разному. Одни соглашались с Альтманом: нужно сделать Microsoft довольной. Компания давала OpenAI деньги и огромные вычислительные мощности. А без этих ресурсов, рассуждали они, невозможно выполнить главную миссию — создать AGI на благо человечества. Другие видели в модели генерации кода огромное экономическое значение. Это идеально соответствовало собственному определению AGI, которое использовала OpenAI: **«высокоавтономные системы, превосходящие человека в большинстве видов экономически ценной работы».** Программирование как раз было такой работой. Но Альтман видел здесь ещё одно преимущество. Если ИИ способен писать код, значит, он может ускорить работу **самой OpenAI**. А в перспективе — возможно, начать помогать компании создавать следующие поколения искусственного интеллекта. Позднее эта идея выльется в проект под названием **AI Scientist** — попытку создать систему, способную автономно заниматься исследованиями в области ИИ.

Для многих исследователей существовал и третий, более научный аргумент. Генерация кода могла стать важной ступенью на пути к следующей модели GPT. Учёные надеялись, что следующая система научится хотя бы в какой-то степени **рассуждать** — способности, которой GPT всё ещё явно не хватало. В более широкой научной среде тем временем продолжался спор между лагерями Хинтона и Маркуса: способно ли глубокое обучение само по себе породить настоящее рассуждение? Исследователи OpenAI предположили: если это вообще возможно, то обучение модели на программном коде должно помочь. Почему? Потому что код — один из крупнейших доступных массивов данных, в которых уже изначально зашиты **структурированные логические закономерности**. Программа строится на причинности, последовательности, условиях, правилах. Если модель научится хорошо работать с кодом, возможно, она перенесёт часть этой способности и на обычный язык. И аргумент снова замыкался на миссии компании: если генерация кода приближает модели к AGI, то разве может быть лучший способ выполнять миссию OpenAI?

Однако даже миллиарды строк кода на GitHub составляли сравнительно небольшой объём по сравнению с данными, на которых обучалась GPT-3. Поэтому команда решила: для лучшего результата модель генерации кода нужно обучать одновременно и на GitHub, и на исходном наборе данных GPT-3. Кроме того, стали искать новые источники. Среди них были: **Stack Overflow** — огромный форум, где программисты задают друг другу вопросы; руководства по программированию; учебники по языкам программирования; другие технические материалы. Возник ключевой технический вопрос: как всё это объединить? Можно было взять уже готовую GPT-3 и дополнительно обучить её на коде — то есть провести **fine-tuning**. А можно было создать новую модель с нуля, сразу смешав старые и новые данные. Команда выбрала дообучение существующей GPT-3. Главная причина была прозаической: **деньги.** Каждый отдельный эксперимент уже стоил около **100 тысяч долларов**, если считать по ценам облачных сервисов Microsoft. Полное обучение новой модели могло обойтись в **десятки миллионов долларов**.

Позднее, когда OpenAI стала создавать модель, которую внутри назовут **GPT-3.5**, подход изменили. Там различные типы данных уже смешивали с самого начала обучения. И внутренние тесты OpenAI действительно показали интересный результат: добавление программного кода улучшало

способности модели решать логические задачи **не только в программировании, но и на обычном английском языке**. Это явление называется **переносом обучения — transfer learning**. Модель осваивает определённую структуру в одной области и затем неожиданно использует её в другой.

Летом 2021 года OpenAI передала GitHub и Microsoft первую рабочую версию модели генерации кода. Она получила название: **Codex.** Но продукт был ещё очень сырым. Модель оказалась слишком большой и слишком медленной. Это означало сразу две проблемы: она была дорогой в эксплуатации, и пользоваться ею было неудобно. Первые совместные работы OpenAI, GitHub и Microsoft быстро породили трения. Никто толком не понимал: кто должен превращать исследовательскую модель в нормальный продукт? OpenAI? GitHub? Неясно было и другое: сколько интеллектуальной собственности OpenAI следует раскрывать коллегам из GitHub? Сотрудники GitHub, в свою очередь, не всегда понимали, насколько можно доверять OpenAI. Затем начались споры о сроках запуска. И ещё более чувствительный вопрос: **кто получит славу?**

В конце концов Мира Мурати сумела договориться о компромиссе. Именно такие ситуации постепенно создавали ей репутацию человека, способного примирять конфликтующие команды разных компаний. Схема была такой: в июне 2021 года Microsoft и GitHub первыми запускают потребительский продукт — **GitHub Copilot.** А через два месяца, в августе, OpenAI выпускает собственную версию Codex через API. Для Microsoft сделка оказалась вполне успешной. GitHub Copilot быстро набирал пользователей. Через два года у него был уже **миллион платных подписчиков**, а ежегодная повторяющаяся выручка превысила **100 миллионов долларов**.

Но для OpenAI этот успех породил совсем другое чувство. В компании всё сильнее укреплялась мысль: **может быть, нам вообще не стоит отдавать свои технологии другим — лучше самим делать продукты для конечного пользователя.** Исследователи OpenAI тяжело работали над моделью. А вся публичная узнаваемость доставалась GitHub и Microsoft. Наблюдать, как две другие компании получают славу за технологию, созданную в значительной степени OpenAI, было неприятно. Кроме того, Microsoft оказалась непростым партнёром. Многие сотрудники OpenAI считали корпорацию слишком бюрократичной. Чтобы Microsoft могла эффективно использовать модели OpenAI, её приходилось постоянно направлять и сопровождать. Но ещё важнее было другое. Если технологии OpenAI доходили до публики через Microsoft, сама OpenAI теряла: контакт с пользователями, данные об их поведении, и, самое главное, **контроль над собственным видением будущего.**

Пока OpenAI сосредоточивала всё больше ресурсов на нескольких ключевых ставках, Сэм Альтман применял похожий принцип и к своим личным проектам. С годами он всё сильнее увлекался так называемыми **hard tech** — технологиями, требующими огромных денег, долгих сроков и решения фундаментальных научных проблем. Он пришёл к убеждению: по-настоящему важные вещи требуют **очень больших ставок и очень длинного горизонта**. Когда Альтман уходил из Y Combinator, его брат Джек спросил на одном мероприятии:

— Если бы ты мог взмахнуть волшебной палочкой и изменить что угодно в мире технологий, стартапов и предпринимательства — что бы ты изменил? Альтман ответил примерно так: «Я бы заставил всех в этой экосистеме мыслить гораздо более длинными временными горизонтами». По его словам, невозможно сделать что-то действительно важное, если человек создаёт компанию, рассчитывая руководить ею четыре-пять лет, или устраивается на работу, заранее предполагая уйти через год-два. Большие вещи требуют десятилетий.

И 2021 год действительно стал переломным в инвестиционной стратегии Альтмана. Раньше он делал много небольших ставок. Теперь решил делать **мало ставок — но огромных**. Одним из проектов стала основанная им ещё в 2019 году компания **Tools for Humanity**. До 2021 года она почти не появлялась в новостях. Но затем начала быстро привлекать финансирование, внимание прессы и

наращивать операции. Главной идеей была попытка создать рабочий механизм для **универсального базового дохода — UBI**. В Кремниевой долине эта концепция давно была популярна: каждому человеку регулярно выплачивается минимальная сумма независимо от того, работает он или нет. Альтман часто говорил об UBI как о возможном спасении в будущем, где искусственный интеллект может вызвать массовые экономические потрясения и вытеснить огромное число работников. Ещё в Y Combinator он организовал крупнейший на тот момент в США эксперимент по изучению безусловного дохода. Для управления проектом появилась некоммерческая организация **OpenResearch**. В течение трёх лет тысяча случайно выбранных малообеспеченных участников из группы в три тысячи человек получала по **1000 долларов в месяц**. Остальные, контрольная группа, получали по 50 долларов. В июле 2024 года OpenResearch опубликовала результаты. Безусловные выплаты помогали людям оплачивать базовые потребности, оказывать помощь другим и чувствовать большую экономическую свободу.

Главным продуктом Tools for Humanity стала криптовалюта **Worldcoin**. Её описывали как «коллективно принадлежащую» валюту, долю стоимости которой в перспективе должен был получить каждый человек. Но здесь возникала проблема: как доказать, что перед вами реальный и уникальный человек? Ответом стала эффектная металлическая сфера размером примерно с шар для боулинга — **Orb.** Устройство сканировало радужную оболочку глаза человека и подтверждало его личность. Только после этого пользователь мог получить свою долю. Основатели утверждали, что подобная система станет особенно важной, когда искусственный интеллект сделает почти невозможным отличить поддельные изображения, видео и личности от настоящих. Но позднее большое расследование журналистов Эйлин Го и Ади Ренальди из *MIT Technology Review* обнаружило вокруг программы сканирования радужки серьёзные проблемы: нарушения приватности, вводящий в заблуждение маркетинг, и возможные нарушения закона. Когда Worldcoin официально запустился в июле 2023 года, разгорелся огромный скандал. Особенно активно людей сканировали в странах Глобального Юга. Тысячи человек вставали в очереди, отдавая компании свои **биометрические данные**, зачастую почти не понимая, для чего именно это делают. Главным стимулом было туманное обещание: **бесплатные деньги.**

В том же 2021 году Альтман сделал две крупнейшие личные инвестиции в своей жизни. **180 миллионов долларов** он вложил в компанию **Retro Biosciences**, занимавшуюся продлением человеческой жизни через клеточное омоложение. Ещё **375 миллионов долларов** — в **Helion Energy**, которая пыталась коммерциализировать термоядерный синтез. Позже Альтман рассказал *MIT Technology Review*: «Я фактически взял всё своё ликвидное состояние и вложил его в эти две компании». И обе технологии он описывал почти тем же языком, каким OpenAI описывала собственную дорожную карту. Сегодня они кажутся почти невозможными. Но если достаточно агрессивно наращивать масштаб — **прорыв может оказаться совсем рядом.**

Ставка на Retro Biosciences отражала давний интерес Альтмана к **долголетию**. Он внимательно следил, например, за исследованиями так называемой «молодой крови» — попытками понять, можно ли обращать некоторые процессы старения вспять с помощью переливания крови молодых организмов. Любопытно, что этой темой интересовался и Питер Тиль, что породило массу публикаций и мемов о его якобы желании вводить себе кровь подростков. Ещё работая в Y Combinator, Альтман внёс **10 тысяч долларов**, чтобы попасть в очередь клиентов спорного стартапа **Nectome**. Идея звучала буквально как сюжет научной фантастики. Компания предлагала: сохранить мозг человека криогенным способом, а затем, возможно через сотни лет, загрузить его содержимое в компьютер — если человечество когда-нибудь научится это делать. Но существовала небольшая проблема. Для качественной консервации мозг должен был быть максимально свежим. Сооснователь компании Роберт Макинтайр сформулировал это журналисту Антонио Регаладо предельно ясно: продукт «на сто процентов смертелен».

Helion была проявлением другой одержимости Альтмана: найти источник **чистой, дешёвой и практически неисчерпаемой энергии**. Он часто говорил, что доступность энергии тесно связана с качеством жизни и экономическим ростом. Но если растущее энергопотребление не перевести на безуглеродные источники, человечество, по его выражению, просто: **«уничтожит планету».** В 2023 году Альтман называл Helion уже не просто инвестицией. Он говорил, что это: «вторая вещь после OpenAI, на которую я трачу очень много времени». Позднее Microsoft заключила соглашение о покупке электроэнергии с первой электростанции Helion. Это произошло уже после третьего инвестиционного раунда Microsoft в OpenAI — на **10 миллиардов долларов**. Helion при этом пообещала запустить первую станцию к **2028 году**. Многие специалисты по энергетике встретили этот срок с удивлением и большим скепсисом.

В 2021 году Альтман перенёс свою страсть к инвестициям и внутрь самой OpenAI. В мае он объявил о создании **OpenAI Startup Fund** — фонда объёмом **100 миллионов долларов** для инвестиций в молодые компании с, по его выражению: «большими идеями о том, как использовать ИИ, чтобы изменить мир». Microsoft снова стала одним из инвесторов. Некоторым наблюдателям решение казалось странным. OpenAI сама почти не зарабатывала денег и при этом уже требовала колоссальных капиталовложений. Зачем ей привлекать дополнительные средства — чтобы потом отдавать их другим компаниям? Одни считали, что Альтман пытался воспроизвести вокруг OpenAI мощный сетевой эффект, который когда-то создал вокруг Y Combinator. Другие видели в этом просто старую привычку. Один человек, работавший с Альтманом, сформулировал так: «Так Сэм и движется по жизни. Заключает сделки».

Альтман любил подчёркивать, что **не владеет долей в OpenAI**. По его словам, это было принципиальным решением: он не хотел, чтобы личная жажда прибыли исказила миссию по созданию безопасного AGI. Его официальная годовая зарплата составляла всего **65 тысяч долларов**, а своё состояние он зарабатывал на других инвестициях. Эта история звучала красиво. И идеально перекликалась с первоначальным объяснением того, почему OpenAI была создана как некоммерческая организация. Но к 2021 году, пишет автор, она уже не отражала всей правды. У Альтмана оставалась значительная доля в Y Combinator. А YC, инвестировавшая в OpenAI **10 миллионов долларов**, потенциально могла получить до **1 миллиарда долларов** возврата. Кроме того, по мере коммерциализации OpenAI многие стартапы Y Combinator и другие компании, в которые Альтман лично вложил деньги, становились клиентами или коммерческими партнёрами OpenAI. Позднее *The Wall Street Journal* подсчитала, что к июню 2024 года совокупное состояние Альтмана во всех его активах составляло как минимум: **2,8 миллиарда долларов.** Создание OpenAI Startup Fund добавляло ещё один слой сложности к рассказу о полной бескорыстности Альтмана. И впоследствии именно эта история сыграет пусть небольшую, но свою роль в его кратковременном отстранении от руководства OpenAI.

Глава 9

Капитализм катастроф Пока OpenAI, ведомая убеждениями Сэма Альтмана, всё быстрее двигалась вперёд, последствия её стратегии становились всё масштабнее. Компания загружала в модели всё более огромные и всё менее очищенные массивы данных. Это и привело к тому самому «сдвигу парадигмы», о котором позже говорил глава Appen Райан Коллн: вместо тщательной фильтрации данных на входе индустрия всё больше переходила к контролю того, что модель выдаёт на выходе. Абстрактный технический язык, однако, скрывал довольно мрачную реальность.

Потому что контролировать эти выходы должны были люди. И именно на них ложилась основная тяжесть новой системы. В 2021 году, одновременно с разработкой следующих поколений моделей, OpenAI начала создавать значительно более совершенный автоматизированный фильтр модерации контента. GPT-3 первоначально появилась в API вообще без фильтрации, что уже привело к скандалу с текстовой платформой Latitude, где модель использовалась для создания сексуального контента с участием детей. Поэтому с моделями, которые позднее назовут GPT-3.5 и GPT-4, компания решила быть осторожнее. OpenAI готовилась распространять свои технологии всё шире, и полностью нефильтрованный продукт рано или поздно мог обернуться серьёзными юридическими, репутационными и практическими проблемами. ChatGPT тогда ещё даже не был задуман. Но впоследствии тот же самый фильтр будет использоваться и вокруг него.

Он должен был работать как защитная оболочка: отслеживать нежелательный контент и удалять его ещё до того, как ответ модели попадёт пользователю. Однако для создания автоматического фильтра OpenAI сначала требовались люди. Очень много людей. Им предстояло вручную просмотреть и классифицировать сотни тысяч примеров именно того контента, который компания хотела не допустить в ответах своих моделей: секса; насилия; жестокого обращения; ненависти; других крайне тяжёлых материалов. После шести месяцев поисков OpenAI нашла подрядчика, казавшегося подходящим для такой работы. Это была аутсорсинговая компания **Sama** — по иронии судьбы её название совпадало с прозвищем Сэма Альтмана. Sama с 2019 года уже занималась модерацией контента для Meta.

OpenAI написала компании письмо и спросила, берётся ли она за проекты с чувствительным и откровенным контентом и как обычно защищает работников от последствий такой работы. Sama дала подробные ответы. В итоге OpenAI подписала с ней четыре контракта общей стоимостью **230 тысяч долларов**. Работа досталась нескольким десяткам сотрудников в Кении. Автор подчёркивает: то, что именно Кения стала местом, где выполнялась одна из наиболее тяжёлых и эксплуататорских работ, связанных с созданием ChatGPT, не было случайностью. Кремниевая долина уже много лет отправляла туда часть своей самой неприятной работы. Кения обладала тем же сочетанием обстоятельств, которое характерно и для других подобных стран:

она бедна; находится в Глобальном Юге; а её правительство остро заинтересовано в иностранных инвестициях. По мнению автора, всё это во многом связано с колониальным наследием страны. Институты защиты трудовых прав здесь развиты слабее, экономика регулярно переживает кризисы, а значит иностранным компаниям легче найти людей, готовых выполнять дешёвую сдельную работу почти в любых условиях. Это неравенство буквально видно на улицах Найроби. В центре города возвышаются стеклянные офисные башни, международные пятизвёздочные гостиницы и дорогие рестораны. Дипломатические кварталы заполнены внушительными зданиями за высокими заборами. В районах, где живут иностранные специалисты, стоят роскошные виллы с зелёными садами. Но стоит уехать на окраины — в Утавалу, Дагоретти-Саут или Эмбакаси, — и город меняется. Сталь и стекло сменяются невысокими бетонными строениями. Здания разбросаны беспорядочно.

Асфальтированные дороги превращаются в грунтовые. Четыре полосы сужаются до узких проездов для мотоциклов и пешеходов. А дальше бетон сменяется рифлёным металлом, из которого буквально друг на друга налеплены самодельные дома и лавки. В этих условиях кенийское правительство охотно принимало технологические компании, искавшие дешёвую рабочую силу. Собственная промышленность страны ограничена. Многие крупнейшие бренды — европейские и

американские. Инфраструктуру, которую когда-то строили британцы, теперь во многом строят китайские компании. Большинство автомобилей — подержанные машины из Японии. Поэтому технологические гиганты воспринимались как источник крайне необходимых рабочих мест. Безработица подпитывает преступность. Мелкие кражи распространены. Люди, чувствующие собственное бессилие, всё меньше доверяют государству.

Во время войны России и Украины, когда в Кении подорожало зерно, распространялись слухи, будто президент намеренно ухудшает положение уже голодающих семей. Звучали даже знакомые по США обвинения: **«Выборы были сфальсифицированы».**

Так Кения постепенно превратилась в один из важнейших центров скрытого труда, поддерживающего работу интернета. Компании вроде Sama — посредники между технологическими гигантами и дешёвой рабочей силой — открывали офисы в Найроби и создавали огромные кадровые резервы для компаний, прежде всего из района Сан-Франциско. На первый взгляд Sama выглядела для OpenAI почти идеальным партнёром. Изначально компания называлась **Samasource**. Она была основана в Сан-Франциско в 2008 году как социальное предприятие, целью которого было давать достойную работу людям из бедных стран и таким образом помогать им выбираться из нищеты. Её основательница Лейла Джана создала подразделения в Индии и Кении и сумела заработать репутацию одного из самых этичных игроков в сфере аутсорсинга. В 2018 году компания превратилась в коммерческую, сократила название до Sama и начала активнее масштабироваться. В 2020 году она получила сертификацию B Corp.

Когда в 2021 году OpenAI задавала Sama вопросы о модерации, компания подробно описала свой опыт работы с тяжёлым контентом, меры по сохранению конфиденциальности и защите данных. Она также подчёркивала, что предоставляет сотрудникам психологическую поддержку. Но за внешним фасадом внутри Sama уже начинались серьёзные проблемы. В январе 2020 года Лейла Джана умерла от редкой формы рака. Ей было всего тридцать семь. Затем началась пандемия. По словам сотрудников, именно тогда в компании усилились управленческие проблемы, хотя представитель Sama эту оценку отрицал. В начале 2022 года кризис стал публичным. Журналист *Time* Билли Перриго опубликовал крупное расследование. Оказалось, что Sama уже выполняла для Meta модерацию Facebook почти по всей Африке к югу от Сахары. Работникам регулярно приходилось смотреть крайне жестокие видеозаписи — самоубийства, обезглавливания и другие сцены насилия.

Многие из них получили тяжёлые психологические травмы. Sama защищала проект, утверждая, что её команда в Восточной Африке согласилась на эту работу после тщательного обсуждения, поскольку хотела, чтобы «африканский контент эффективно проверялся самими африканцами». Позднее почти двести работников подали несколько исков против Sama и Meta. Они заявляли о травмирующих условиях труда и незаконных увольнениях тех, кто пытался объединяться ради повышения зарплаты и улучшения условий. Sama эти обвинения отвергала. Именно на этом фоне в конце 2021 года начался проект OpenAI. Внутри него использовались кодовые названия **PBJ1, PBJ2, PBJ3 и PBJ4**.

Кенийские сотрудники Sama снова оказались перед крайне тяжёлым контентом. Средняя оплата составляла примерно от **1,46 до 3,74 доллара в час**. При этом сами работники не знали, кто является конечным заказчиком и зачем выполняется проект. Условия неразглашения не позволяли им этого выяснить — обычная практика в индустрии разметки данных. Они знали только то, что видели перед собой. Сотни тысяч чудовищных текстовых описаний, которые нужно было читать и распределять по категориям: это просто насилие или крайне графическое насилие?

домогательство или язык ненависти? сексуальное насилие над детьми или зоофилия? Постепенно эта работа начала ломать людей. И последствия выходили далеко за пределы каждого отдельного сотрудника: страдали семьи и люди, которые от него зависели.

Лишь после появления ChatGPT многие из них начнут понимать, ради какой технологии они заплатили собственным душевным спокойствием. В мае 2023 года автор встретилась в Найроби с четырьмя такими работниками. Они согласились рассказать свои истории для материала на первой полосе *The Wall Street Journal*.

Одного из них звали **Мофат Окиньи**. Он работал в команде, занимавшейся сексуальным контентом. Позднее выяснится, что проект, разрушивший его психологическое состояние и отношения с близкими, помогал создавать технологию, которая затем сама начала подрывать экономические возможности его собственного брата.

Ещё задолго до нынешней эпохи генеративного ИИ исследователи начали замечать скрытую рабочую силу, без которой индустрия искусственного интеллекта вообще не могла бы существовать. Антрополог Microsoft Мэри Л. Грей и специалист по вычислительным социальным наукам Сиддхарт Сури были среди первых, кто подробно рассказал миру о людях, подобных кенийским разметчикам. В 2019 году они опубликовали книгу *Ghost Work* — «Призрачный труд».

Она стала результатом пяти лет полевых исследований и раскрывала огромную скрытую сеть мелких цифровых работ, на которой фактически держалась Кремниевая долина. Оказалось, что миллиардные оценки технологических гигантов создавались не только инженерами с шестизначными зарплатами в модных офисах. Не менее необходимыми были люди — часто в странах Глобального Юга, — которые за гроши вручную размечали бесконечные массивы данных.

Хороший пример — беспилотные автомобили. Чтобы машина могла самостоятельно ехать по правильной полосе, реагировать на опасного водителя и останавливаться перед школьниками на переходе, её программное обеспечение использует множество моделей глубокого обучения. Одни из них должны распознавать объекты: разметку полос; дорожные знаки; светофоры; машины; деревья; пешеходов.

Чтобы научить модели всему этому, компании отправляют автомобили с многочисленными камерами на дороги и записывают огромные объёмы видео. Но сами видеозаписи недостаточны. Кто-то должен буквально кадр за кадром обвести каждый объект. Иногда — с невероятной точностью: изгиб руки человека, держащего руль велосипеда, или собаку, наполовину высунувшуюся из окна автомобиля. После этого каждому объекту присваивают метку: «велосипед»; «автомобиль»; «животное»; «человек». Всё это делают люди. И с точки зрения компании действует простое правило:

чем дешевле они это делают, тем лучше. Грей и Сури много изучали платформу **Amazon Mechanical Turk**. Она появилась ещё до бума глубокого обучения и долгое время служила универсальным рынком дешёвого цифрового труда. Любая компания могла разместить там небольшую работу и найти человека, готового выполнить её за небольшую сумму. К моменту выхода *Ghost Work* индустрия уже начала переходить к новому этапу.

Появлялись платформы, специализирующиеся именно на разметке данных для искусственного интеллекта. Автор книги начала исследовать эту новую глобальную рабочую силу. Она разговаривала с десятками работников, ездила к ним домой, ужинала с их семьями, пыталась понять не только общие экономические закономерности, но и то, как эта система ощущается в обычной повседневной жизни. И постепенно стало ясно: эксплуатация работников в эпоху генеративного ИИ возникла не внезапно. Она была встроена в отрасль задолго до ChatGPT.

До генеративного ИИ главным двигателем индустрии разметки данных были беспилотные автомобили. Volkswagen, BMW и другие старые автомобильные гиганты испугались конкуренции со стороны Tesla и Uber и начали создавать собственные подразделения беспилотного транспорта. В эту гонку хлынули миллиарды долларов. А вместе с ними резко вырос спрос на размеченные данные. Mechanical Turk уже не справлялся. Он был слишком универсальным и примитивным. Интерфейс выглядел так, будто застрял где-то в середине 2000-х: компания загружала набор данных, писала простые инструкции, назначала цену. Имена работников заменялись случайными комбинациями букв и цифр. Рядом было две кнопки: выдать бонус;

или выгнать человека с проекта. Но для беспилотных автомобилей требовалась совсем другая точность. Одна небрежно размеченная машина или пропущенный пешеход могли в реальном мире означать разницу между жизнью и смертью. Работников нужно было обучать. Инструкции должны были быть подробными. Требовались проверки, обратная связь и повторная разметка. Поэтому на место Mechanical Turk пришли новые компании:

Scale AI, Hive, Mighty AI, Appen. И тут произошло нечто неожиданное. На этих платформах внезапно стали массово регистрироваться люди из Венесуэлы. Причина была трагической. Как раз тогда, когда автомобильные гиганты вкладывали миллиарды в беспилотники, а компании по разметке desperately искали дешёвых работников, Венесуэла погружалась в один из тяжелейших экономических кризисов мирного времени за последние полвека. Некогда это была одна из богатейших стран Латинской Америки, обладавшая крупнейшими разведанными запасами нефти в мире. Но с 2016 года гиперинфляция вышла из-под контроля. Безработица резко выросла. Преступность усилилась. Сбережения семей превращались практически в ничто. С конца 2017 по 2019 год санкции администрации Дональда Трампа, введённые в ответ на авторитарную политику Николаса Мадуро, нанесли экономике ещё один тяжёлый удар.

Гиперинфляция достигла немыслимых **10 миллионов процентов**. Люди с университетскими дипломами и бывшими хорошо оплачиваемыми профессиями теперь часами стояли в очередях ради небольших порций риса и муки. В этой катастрофе сотни тысяч венесуэльцев нашли один из немногих способов заработать: разметку данных через интернет. К середине 2018 года в некоторых компаниях венесуэльцы составляли до **75 процентов всей рабочей силы**. Иногда работала вся семья. Родители и дети по очереди садились за один компьютер. Жёны брали на себя домашние обязанности, чтобы мужья могли чуть дольше не отрываться от оплачиваемых заданий. Почему именно Венесуэла? На первый взгляд страна была далеко не идеальной. Большинство компаний находились в Сан-Франциско и Сиэтле. Возникал языковой барьер. Но экономическое отчаяние имело решающее значение. Люди были готовы работать за невероятно низкую оплату. А значит, платформы могли предлагать своим клиентам невероятно дешёвые услуги. Один исследователь индустрии назвал это «странным совпадением».

Но автор увидела в нём нечто гораздо более тревожное. Экономический крах создал почти идеальные условия для дешёвого цифрового труда: население было хорошо образовано; интернет был достаточно развит; люди отчаянно нуждались в любой работе. Венесуэла была не единственной страной, где могло возникнуть такое сочетание. Интернет становился доступнее всё большему числу людей. Климатические потрясения усиливались. Геополитическая нестабильность росла. Поэтому исследователь Флориан Александер Шмидт предположил: **«Вполне вероятно, появится ещё одна Венесуэла».**

Для автора это породило неприятный вопрос. Если в следующий раз подобная индустрия сознательно отправится искать работников в страну, переживающую кризис, можно ли всё ещё называть это совпадением? Или кризис сам превратится в бизнес-модель? Спустя несколько лет, пишет она, именно это и произошло. И здесь возникает одна из самых мрачных параллелей между старыми империями и новыми «империями искусственного интеллекта»: быстро накапливать богатство особенно легко, когда можно почти ничего не платить огромному числу людей, чьим трудом это богатство создаётся.

В декабре 2021 года автор отправилась через горные районы Колумбии, чтобы лучше понять жизнь человека, который в условиях кризиса оказался в индустрии разметки данных. Поехать в Венесуэлу тогда было невозможно из-за ограничений на поездки. Но в соседней Колумбии к тому времени жили почти два миллиона венесуэльских беженцев — примерно треть всех людей, покинувших страну из-за экономической катастрофы.

Профессор Йельского университета Хулиан Посада познакомил автора с одной из них — **Оскариной Вероникой Фуэнтес Анайей**. Фуэнтес продолжала работать в сфере разметки данных уже после того, как бежала из Венесуэлы. Именно она впервые по-настоящему показала автору, как устроена эта работа изнутри: как человек постепенно перестраивает вокруг платформы всю свою жизнь — и как сама платформа при этом относится к нему как к чему-то полностью заменяемому.

Они сидели рядом в гостиной маленькой квартиры, которую Фуэнтес делила примерно с шестью родственниками, а она одну за другой открывала страницы Appen. Задания были самыми разными. Одни касались интернет-магазинов: к какой категории отнести товар — «одежда» или «аксессуары»? Другие — модерации социальных сетей: есть ли в этом видео сцены преступления или нарушения

прав человека? Когда задание было на английском, Фуэнтес переводила текст на родной испанский через Google Translate.

За каждое выполненное задание сумма её заработка увеличивалась буквально на несколько центов. Чтобы вывести деньги, нужно было накопить не меньше **десяти долларов**. Когда Фуэнтес только начала работать на платформе, это не составляло особой проблемы. Теперь на эти десять долларов иногда уходили недели. И порог минимальной выплаты начинал казаться почти жестоким арбитром, от которого зависело, хватит ли ей денег на продукты.

Для тех, кто оставался в самой Венесуэле, получить заработанные деньги было ещё сложнее. Большинство глобальных платёжных систем, включая PayPal, не позволяли переводить средства в Венесуэлу. А магазины внутри страны часто не принимали те способы оплаты, которые всё-таки работали. Поэтому онлайн-заработок приходилось превращать в наличные. Деньги приходили в долларах США, а для обычных покупок были нужны венесуэльские боливары. Возник огромный чёрный рынок обмена — с мошенничеством, высокими комиссиями и постоянным риском потерять и без того небольшую сумму.

Отношения Фуэнтес с Appen были сложными. Она никогда не мечтала о такой работе. Просто целая цепочка обстоятельств, которые она не могла контролировать, превратила платформу одновременно и в спасательный круг, и в источник постоянного давления.

Аккаунт на Appen она открыла ещё во время магистратуры по инженерной специальности — просто чтобы немного подзаработать. Она была способной, трудолюбивой и изобретательной. Если бы её страна не рухнула экономически, такая студентка, вероятно, без особых проблем получила бы стабильную работу в государственной нефтяной компании. Но страна рухнула. И Фуэнтес пришлось приспосабливаться.

Она тщательно спланировала бегство в Колумбию вместе с мужем. В каком-то смысле ей повезло. По праву рождения Фуэнтес могла получить колумбийский паспорт. У многих других венесуэльцев, спасавшихся от кризиса, документов не было. Её родители сами были колумбийцами. Поколением раньше они бежали в противоположном направлении — из Колумбии в Венесуэлу, спасаясь уже от другой волны насилия и политической нестабильности. Теперь история повернула вспять.

Одна страна переживала кризис — семья бежала в другую. Через поколение кризис приходил уже туда. Автор видит в этом знакомую картину: бедствия наслаиваются друг на друга, переходят через границы и заставляют семьи поколениями жить в режиме выживания.

Благодаря паспорту Фуэнтес ещё из Венесуэлы договорилась снять квартиру у знакомого в Колумбии. Для аренды требовались два поручителя, владевшие недвижимостью. Будущий хозяин квартиры обещал помочь их найти. В начале 2019 года Фуэнтес и её муж пересекли границу. Денег у них было примерно на неделю еды.

Когда они добрались до квартиры, выяснилось, что там уже живёт другая венесуэльская пара, которой хозяин тоже пообещал это жильё. Выбора не было. Обе пары поселились под одной крышей, при этом каждая боялась, что другая в конечном счёте вытеснит её из квартиры. Позже вторая пара уехала. Но проблемы только начинались. Фуэнтес устроилась в местный колл-центр. Её муж пока не имел разрешения на работу. Не успел он его получить, как колл-центр объявил о закрытии.

И именно в этот момент у Фуэнтес появились признаки серьёзного заболевания. Она их проигнорировала. Ей казалось, что сейчас самое главное — отработать как можно больше часов, пока работодатель ещё существует.

Позже врач сказал ей, что если бы она пришла в больницу чуть позже, то, вероятно, умерла бы. Коллега заметил, насколько плохо она выглядит, и отвёз её в больницу. Вскоре после этого у Фуэнтес начались судороги.

Её пульс остановился примерно на минуту. Диагноз — тяжёлая форма диабета. Ей сразу назначили пять ежедневных инъекций инсулина. Несколько недель она страдала от сильной боли и приступов временной слепоты.

Когда состояние стабилизировалось, осталась постоянная изматывающая усталость. Она уже не могла надолго выходить из дома. Но даже тогда главным вопросом для неё оставались деньги. С

хроническим заболеванием ездить далеко в офис стало небезопасно. И тогда Фуэнтес снова достала ноутбук. И снова вошла в Appen.

Фуэнтес совершенно не понимала, по какой логике задания появляются в её очереди. Ясно было только одно: если хочешь продолжать получать работу, нужно постоянно демонстрировать высокую точность и скорость. Тогдашний технический директор Appen Уилсон Панг объяснял автору, что платформа распределяла задания алгоритмически. Учитывалось сразу несколько факторов: местоположение; общая точность; скорость; типы задач, с которыми работник раньше справлялся особенно хорошо. Но для работников всё выглядело как загадочная система с невидимыми правилами.

В Telegram- и Discord-группах венесуэльцы делились догадками, пытаясь вместе разгадать алгоритм. Они выяснили, например, что если использовать VPN и заставить систему думать, будто ты находишься в США, можно получить более высокооплачиваемые задания. Но это было крайне рискованно. Appen отслеживала такие нарушения. Аккаунт могли закрыть. А закрытие аккаунта для работника означало катастрофу. Все невыведенные деньги исчезали.

Приходилось создавать новый профиль и начинать с самого низа — с самых дешёвых заданий. А всё чаще — вообще без заданий. Существовали и другие странные правила. Быстро выполнить задание — хорошо. Слишком быстро — плохо. Если система считала, что человек выполнил работу подозрительно быстро, она могла решить, что используется бот, и просто не заплатить. Иногда задания приходили практически без инструкций. Иногда сама платформа работала с ошибками и ничего нормально не загружала.

Среди венесуэльских работников были бывшие программисты. Они начали писать расширения для браузера и бесплатно делиться ими с другими. Одно специально задерживало отправку каждого задания, чтобы алгоритм не решил, что работает бот. Другое автоматически обновляло очередь каждую секунду. Третье включало сигнал тревоги, когда появлялось новое задание.

Благодаря этому работник хотя бы мог отойти в туалет или приготовить еду, не боясь упустить работу. Люди помогали друг другу. А сама платформа одновременно заставляла их конкурировать между собой. Работа распределялась по принципу: **кто первый успел — того и задание.** Оно оставалось доступным лишь до тех пор, пока нужное количество работников не успевало его забрать. Поначалу этот промежуток мог длиться днями. Потом часами. А затем — секундами. На платформу приходило всё больше людей, особенно венесуэльцев, загнанных кризисом в отчаянное положение. И все они боролись буквально за крохи работы. Постепенно непредсказуемый ритм платформы начал полностью управлять жизнью Фуэнтес.

Однажды она вышла прогуляться. И тут появилось задание, за которое можно было получить несколько сотен долларов — сумму, достаточную примерно на месяц жизни. Она бросилась домой бегом. Но когда добралась до компьютера, задание уже разобрали другие. После этого случая Фуэнтес почти перестала выходить из дома в будни. Даже по выходным позволяла себе выбраться максимум на полчаса — она заметила, что именно тогда задания появляются реже.

Спала плохо. Её мучила мысль, что выгодная работа может появиться посреди ночи. Перед сном она выкручивала громкость компьютера на максимум.

Если ночью появлялось задание, специальное расширение должно было разбудить её сигналом. И всё же, несмотря на стресс и постоянное напряжение, Фуэнтес не могла представить, что однажды просто уйдёт с Appen. Она боялась противоположного: что Appen сама перестанет давать ей задания. Когда в её жизни рушилось всё остальное, эта платформа спасла её. Когда-то заработки были даже настолько хороши, что она смогла купить новый ноутбук — и довольно быстро окупить его стоимость. Когда дела шли хорошо, они шли **очень хорошо**. А когда плохо — она всё равно оставалась привязана к платформе в надежде, что прежние времена вернутся. Автор говорит, что история Фуэнтес научила её двум вещам. Первая была тяжёлой.

Даже если Фуэнтес захотела бы отказаться от платформы, сделать это было бы почти невозможно. Её биография — беженка, ребёнок семьи, поколениями переживающей нестабильность, человек с хронической болезнью — была вовсе не исключительной среди подобных работников. Бедность, пишет автор, — это не просто отсутствие денег. Она проникает во все области жизни. Она накапливает

долги, которые нельзя измерить только валютой: нарушенный сон; ухудшение здоровья; падающая самооценка; и, главное, утрата контроля над собственной жизнью.

Но была и вторая, более обнадёживающая истина. Фуэнтес не ненавидела саму работу. Ей не нравилось **то, как эта работа была организована**. А значит, проблема в принципе решаема. Когда автор спросила её, что она хотела бы изменить, список оказался удивительно простым. Фуэнтес хотела, чтобы Appen стала обычным работодателем. Чтобы был: постоянный трудовой договор; руководитель, с которым можно поговорить; стабильная зарплата; медицинская страховка. Ей и другим работникам, говорила она, нужна была прежде всего **уверенность в завтрашнем дне**.

И ещё одно: они хотели, чтобы компания, ради которой они так много работают, хотя бы знала, что они существуют. Исследователи труда, опрашивавшие работников по всему миру, приходили примерно к тем же выводам. Проект Fairwork при Оксфордском институте интернета сформулировал минимальные условия достойного цифрового труда:

работникам следует платить прожиточную заработную плату; у них должны быть регулярные смены; оплачиваемый больничный; понятный договор; возможность общаться с руководством; и право объединяться в профсоюзы без страха наказания. Некоторые компании в индустрии разметки данных действительно пытались следовать таким принципам. Но у них была одна большая проблема: **цена.** Компании, которые платили людям лучше, проигрывали тендеры тем, кто этого не делал. Если отрасль в целом не устанавливает минимальные стандарты, гонка вниз становится практически неизбежной.

Одной из компаний, особенно успешно освоивших эту модель, стала **Scale AI**. Её в 2016 году основал девятнадцатилетний бывший студент MIT Александер Ван. С самого начала стратегия Scale частично строилась на простой цели: дать клиентам высококачественную специализированную услугу по максимально низкой цене. Один бывший сотрудник, отвечавший за расширение рабочей силы, описал задачу предельно откровенно: **«Как найти самых хороших людей за самые маленькие деньги?»** Scale быстро получила крупных клиентов: Lyft; Apple; Toyota; Airbnb. Если работники Mechanical Turk в основном находились в США и Индии, Scale сначала сосредоточилась на Кении и Филиппинах. Обе страны были бывшими колониями, где широко распространён английский язык и давно существовала индустрия колл-центров и удалённой работы для американских компаний. Команды Scale специально искали места, где сочетались три фактора:

хорошее образование; нормальный интернет; и бедность, заставляющая людей много работать за очень небольшую плату. Но внутри компании существовало и более благородное объяснение. Сотрудники говорили себе, что именно таким людям экономическая возможность нужна больше всего. Мол, они не только зарабатывают, но и становятся счастливее. Один инвестор Scale выразил эту идею так: если выбирать между тяжёлой физической работой и разметкой данных в интернет-кафе с кондиционером, второй вариант всё же лучше. Но когда Scale запустила платформу **Remotasks** и увидела огромный наплыв пользователей из Венесуэлы, страна быстро стала одним из главных направлений набора персонала. Бывший сотрудник объяснил почему предельно просто: **«Они самые дешёвые на рынке».** В 2019 году компания начала активную кампанию по привлечению венесуэльцев. Использовались реферальные коды. Реклама в социальных сетях. Видео со стоковыми кадрами стопок американских долларов. В следующем году Scale даже создала специальную венесуэльскую страницу Remotasks и запустила программу **Remotasks Plus**. Её рекламировали как помощь людям, переживающим исторический кризис. Участникам обещали:

обучение новым навыкам; карьерный рост; стабильные часы; почасовую оплату; более высокий заработок. Началась пандемия, ещё сильнее ударившая по Венесуэле. Люди массово пошли в Remotasks Plus. А конкуренты Scale начали терять позиции. Но как только Scale укрепила своё положение на рынке, обещания начали исчезать. Автор вместе с венесуэльской журналисткой Андреа Паолой Эрнандес изучила историю программы. Они обнаружили открытую таблицу, оставленную компанией в интернете. Из неё следовало, что заработки начали падать уже через несколько недель после запуска. Работники, которые первоначально получали примерно **40 долларов в неделю**, вскоре

стали зарабатывать меньше **шести долларов**. А некоторые — вообще ничего. В апреле 2021 года Scale полностью закрыла Remotasks Plus и вернулась к обычной модели:

отдельные задания; никаких гарантированных часов; никакого стабильного заработка. Внутри Scale программу считали экспериментом. Компания предполагала, что платить людям по часам будет проще, чем за каждое выполненное задание. Но возникла проблема: Scale не могла надёжно проверить, сколько часов люди действительно работали. Сотрудники подозревали, что некоторые пользователи завышают время. Компания начала вводить всё новые формы наблюдения. Но проблему это не решило. В конце концов программу закрыли, чтобы остановить денежные потери. При этом более **85 процентов работников** всё равно остались на платформе.

Представитель Scale приводил это как доказательство «устойчивого интереса и вовлечённости». Но у этих работников фактически не было другого выхода. Через семь месяцев после закрытия Plus заработки упали ещё ниже. Журналистка Эрнандес сама зарегистрировалась на платформе. За два часа — включая обучение и выполнение двадцати заданий — она заработала: **11 центов.** Тогдашний руководитель операций Scale Мэтт Парк заявил, что средний заработок венесуэльцев на платформе составлял чуть более **90 центов в час**.

И добавил: «Remotasks стремится платить справедливую заработную плату в каждом регионе, где мы работаем». Многие венесуэльцы, которые начинали жаловаться, лишались доступа к платформе. Компьютерный инженер Рикардо Хаггинс пришёл в Remotasks после катастрофического общенационального отключения электричества, продолжавшегося неделю. Ему нужно было содержать жену и детей.

По его словам, аккаунт закрыли после того, как он начал задавать слишком много вопросов в Discord-сообществе. Позднее он сформулировал своё впечатление так: **«Я понял, что их подход — выжать из каждого пользователя всё возможное, а потом выбросить его и привести нового».** И новые пользователи действительно приходили. К середине 2021 года, когда венесуэльцы выгорали и покидали платформу, Scale уже искала десятки тысяч работников в других странах, чьи экономики серьёзно пострадали во время пандемии. Компания нуждалась не только в дешёвом труде. Ей требовались носители языков, ценных для клиентов: английского; французского;

итальянского; немецкого; китайского; японского; испанского. Франкоязычных работников искали, например, в бывших французских колониях Африки. Носителей китайского — в странах Юго-Восточной Азии с крупной китайской диаспорой. И снова применялась схема, опробованная в Венесуэле. В новом регионе сначала предлагали привлекательный заработок. Люди приходили. Платформа закреплялась. А затем выплаты уменьшались. В офисе Scale это обсуждали как **«оптимизацию»** и **«инновации»**. Для работников же изменение нескольких цифр в системе означало разрушенный семейный бюджет и потерянный источник существования. Группа из восьми работников в Северной Африке рассказала автору, что за несколько месяцев их оплату сократили более чем на треть.

У одного человека система даже показала отрицательную ожидаемую выплату — словно он сам оказался должен Scale деньги. Когда работники попытались организованно возражать, компания пригрозила блокировать тех, кто участвует в «революциях и протестах». Почти все, кто разговаривал с автором, впоследствии лишились доступа к платформе. Scale заявила, что не блокирует людей за жалобы на оплату, а делает это только за нарушение правил сообщества. Были проблемы и с самими выплатами.

Система работала с ошибками, и люди порой просто не могли вывести заработанное. Некоторые штатные сотрудники Scale, тесно взаимодействовавшие с разметчиками, пытались отстаивать для них лучшие условия, зарплаты и хотя бы гарантии получения уже заработанных денег. Многие в конце концов сами уходили — выгорев или оказавшись вытесненными из компании. Позднее Scale заявила, что значительно улучшила стабильность платформы.

Доминирование Scale одновременно создавало проблему для компаний, которые пытались относиться к работникам иначе. Например, **CloudFactory**, работающая в Кении и Непале, предоставляла людям обычные трудовые договоры и стабильные рабочие часы. Её модель

соответствовала принципам Fairwork. Компания убеждала заказчиков: да, такой труд дороже, зато качество выше.

Люди хорошо обучены. Со временем приобретают опыт. Лучшие получают повышение. Многие кенийские работники, с которыми разговаривала автор, считали CloudFactory одним из лучших работодателей в отрасли. Иногда такая аргументация работала. Некоторые клиенты специально выбирали CloudFactory именно из-за её репутации хорошего работодателя. Но когда во время пандемии бюджеты сократились, многие снова вернулись к более дешёвым вариантам. CloudFactory пришлось увольнять сотрудников. По словам работников Sama, похожее конкурентное давление постепенно ухудшало и условия там.

Когда-то место в Sama считалось даже более престижным, чем в CloudFactory. Затем умерла Лейла Джана. Пришла пандемия. Клиенты начали уходить к Scale и другим более дешёвым конкурентам. По мнению сотрудников, именно эта финансовая ситуация и подтолкнула Sama согласиться на тяжёлый проект OpenAI по модерации контента. Компания оказалась в трудном положении. Практически так же, как сами венесуэльские работники, вынужденные соглашаться на всё менее выгодную работу.

Дальше начинается одна из центральных историй главы — история **Мофата Окиньи**, кенийского работника, которому предстояло вручную просматривать материалы для будущей системы безопасности OpenAI. Он вырос в деревне на острове на западе Кении, посреди озера Виктория. До Найроби оттуда — восемь часов на автобусе и ещё два часа на лодке.

Медицинская помощь была далеко. Внезапная серьёзная болезнь нередко означала смерть. Семья была бедной. Но в детстве Мофат и его братья и сёстры почти не думали о бедности. Гораздо сильнее их занимали рассказы предков. По преданию, народ луо, к которому принадлежала семья Окиньи, когда-то пришёл из Израиля. Владение строительством лодок и речной навигацией позволило им двигаться на юг вдоль Нила и расселиться в западной Кении, Уганде и Танзании.

Окиньи рассказывал это автору почти шёпотом, словно открывал тайну: **«Луо — не кенийцы. Мы израильтяне, живущие в Кении. Но без луо не было бы и Кении».**

А затем с улыбкой добавил: **«Барак Обама — луо. Луо — очень умный народ».** К двадцати восьми годам бедность уже занимала в его мыслях гораздо больше места. Нужно было платить за жильё. Покупать еду. Оплачивать школу племяннице. Государственное образование в Кении не было для семьи полностью бесплатным. Любая работа воспринималась как удача. В 2021 году Всемирный банк оценивал, что более четверти населения страны жило менее чем на **2,15 доллара в день**.

Поэтому звонок из Sama в ноябре 2021 года казался Мофату почти чудом. В компанию он пришёл ещё в 2019 году, откликнувшись на вакансию «обучение ИИ». Первые два года его работа отражала тогдашнее развитие индустрии: он занимался разметкой изображений для компьютерного зрения, в том числе для беспилотных автомобилей. Новый проект должен был стать его первым опытом работы с генеративным ИИ. Но тогда он этого ещё не знал.

Перед началом новой работы менеджеры Sama провели с Мофатом так называемый **тест на психологическую устойчивость**. Ему дали несколько тревожных текстов и попросили классифицировать их по инструкции. Он справился отлично. После этого ему предложили присоединиться к новой команде, где работа, как ему объяснили, будет похожа на модерацию контента. Раньше Мофат этим не занимался, но материалы в тесте показались ему терпимыми. Отказываться от работы посреди пандемии было бы нелепо. К тому же он думал о будущем. Он жил в Пайплайне — шумном, бедном районе на юго-востоке Найроби, плотно застроенном многоквартирными домами, с круглосуточными уличными торговцами и постоянной суетой молодых людей, пытающихся вырваться в лучшую жизнь. Мофату казалось, что и он движется именно туда. Недавно он познакомился с девушкой по имени Синтия. Впервые в жизни рядом с ней он начал всерьёз представлять себе семью.

Лишь после того, как он согласился на новый проект, стало ясно: реальные материалы будут гораздо тяжелее тех, что использовались в тесте. OpenAI разделила работу на несколько потоков. Один был посвящён сексуальному контенту. Другой — насилию, языку ненависти и самоповреждениям. В феврале 2022 года насилие выделили в отдельное направление. В каждом

потоке группа работников Sama — их называли **агентами** — читала тексты и классифицировала их по инструкциям OpenAI. Затем меньшая группа **аналитиков качества** проверяла их решения, прежде чем результаты отправлялись обратно заказчику.

Мофат стал аналитиком качества в команде сексуального контента. По контракту он должен был просматривать около **15 тысяч фрагментов в месяц**. OpenAI делила сексуальный текстовый контент на пять категорий. На самом тяжёлом уровне находились описания сексуального насилия над детьми — любое упоминание сексуальной активности с участием человека младше восемнадцати лет. Следом шли описания сексуального контента, который в реальной жизни мог бы быть незаконным в США: инцест; зоофилия;

изнасилование; сексуальная эксплуатация; сексуальное рабство. Некоторые материалы собирались из самых тёмных уголков интернета — например, с эротических сайтов с фантазиями об изнасиловании или из сообществ, посвящённых самоповреждениям. Другие тексты создавал сам искусственный интеллект. Исследователи OpenAI специально просили языковую модель генерировать крайне тяжёлые сценарии, чтобы потом на этих примерах обучать систему защиты.

Именно здесь эта работа отличалась от обычной модерации. Модератор Facebook рассматривает реальный пользовательский пост и решает, можно ли оставить его на платформе. Мофат и его коллеги делали другое. Они размечали примеры, чтобы научить фильтр OpenAI **заранее не допускать**, чтобы языковая модель сама производила подобный материал. То есть для создания достаточно широкого набора примеров некоторые из худших сценариев специально придумывала сама система. Машину учили не создавать чудовищные вещи, показывая ей огромное количество именно таких вещей.

Поначалу тексты были короткими — одно-два предложения. Мофат старался просто отделять работу от личной жизни. Тем временем его отношения с Синтией развивались стремительно. Он говорил брату Альберту, что нашёл любовь всей своей жизни. У Синтии уже была маленькая дочь от предыдущих отношений, и Мофат относился к ней как к собственной. В начале 2022 года они переехали из Пайплайна в Утавалу — более спокойный жилой район на востоке Найроби. Никаких официальных документов они не оформляли, но по их традиции совместная жизнь означала практически брак. Они называли друг друга мужем и женой. А работа тем временем становилась всё тяжелее.

График сделался непредсказуемым. То вечерние смены. То выходные. Тексты становились длиннее — иногда по пять-шесть абзацев. И всё более подробными. Содержимое начинало преследовать работников не только своей жестокостью, но и мучительной конкретностью. Вокруг Мофата коллеги, особенно женщины, всё чаще не выдерживали.

Люди просили больничные, семейные отпуска, находили любые причины, чтобы не приходить на работу. Sama предоставляла бесплатные психологические консультации. Но многим этого было недостаточно. Сеансы часто проходили в группах, где трудно было открыто говорить о личных переживаниях. К тому же, по ощущениям работников, сами психологи не всегда понимали, с каким именно материалом приходится сталкиваться сотрудникам. Многие вообще боялись обращаться за помощью. Признаться, что тебе плохо, означало признаться, что работаешь уже не так эффективно. А если работаешь хуже — тебя могут заменить. Представитель Sama позднее заявлял, что ни Мофат, ни другие сотрудники не сообщали компании о проблемах с доступом к психологической помощи и что руководство узнало об этом только из прессы.

Мофат пытался просто держаться. Но чувствовал, что начинает разваливаться изнутри. Прочитанные тексты словно впивались в сознание. Они преследовали его дома. Приходили во сне. Возвращались образами, от которых невозможно было избавиться. Ему казалось, что от прежнего человека осталась только оболочка. Он отдалился от друзей. Стал холоднее к приёмной дочери.

Перестал быть близок с женой. В марте 2022 года руководство Sama собрало работников и объявило: контракт с OpenAI прекращается. Часть сотрудников, в том числе Мофата, переведут на другие проекты, не связанные с модерацией. Другие останутся без работы. Многие считали, что резкое решение было связано с расследованием *Time* о работе Sama для Meta. После того как сотрудники начали говорить с журналистами, компания оказалась в центре серьёзного репутационного скандала

и прекратила другие проекты по модерации. Sama давала другое объяснение. По словам её представителя, проект OpenAI с самого начала был пилотным, а контракт прекратили потому, что OpenAI начала присылать для разметки изображения, выходившие за согласованный объём работ. Sama в итоге не получила от OpenAI всю сумму в 230 тысяч долларов.

Но исчезновение самой работы не вернуло Мофату прежнюю жизнь. Его психологическое состояние продолжало ухудшаться. Он страдал бессонницей. Проваливался то в тревогу, то в депрессию. Счастливый период отношений с Синтией закончился. Она требовала объяснить, что происходит. А он не знал, как сказать. Как объяснить любимому человеку, что каждый день ты часами читаешь тексты о сексуальных извращениях и насилии?

Он понимал, что его молчание, должно быть, доводит её до отчаяния. Синтия говорила: ты изменился; ты больше не выполняешь обещаний; ты больше не любишь мою дочь. Мофат снова попытался обратиться за психологической помощью — теперь к частному специалисту.

Одна консультация стоила **1500 кенийских шиллингов**, примерно 13 долларов по курсу 2022 года. То есть больше его дневного заработка. Полный курс лечения, сказал врач, обойдётся примерно в **30 тысяч шиллингов — около 250 долларов**. Практически месячная зарплата. Мофат оплатил одну консультацию. И больше не вернулся.

В ноябре он наконец нашёл новую работу. К счастью, уже не связанную с модерацией контента. Он устроился в службу поддержки клиентов у одного из конкурентов Sama и начал ездить в офис в деловом центре Найроби. Мофат надеялся, что жизнь наконец вернётся в нормальное русло.

Через неделю после начала работы он возвращался домой, когда Синтия написала ему: купи рыбу на ужин. Он купил три куска. Один себе. Один ей. Один дочери. Но когда он пришёл домой, сразу понял: что-то не так. Ни Синтии, ни девочки не было. Их вещей тоже. Из коротких сообщений он понял:

они ушли. И не вернутся. Мофат вспоминал её слова: **«Ты изменился. Ты уже не тот человек, за которого я вышла замуж. Я больше тебя не понимаю».** Его брат Альберт тогда жил в Момбасе, более чем в восьми часах езды от Найроби. Он изучал английскую литературу и преподавал её в школе.

В свободное время писал стихи. В течение многих месяцев он и сам видел, как меняется Мофат, — в коротких эпизодах во время регулярных видеозвонков. Когда брат позвонил ему после ухода Синтии, Альберт сначала не понял. — Мой дом пуст, — сказал Мофат. Альберт решил, что его ограбили. Потом понял, что на самом деле произошло.

И решил немедленно ехать. Он сообщил школе, что увольняется, собрал вещи и переехал к брату — в ту самую квартиру в Утавале, где раньше жили Мофат и Синтия. Для Альберта это решение тоже имело цену.

В Найроби он не смог найти постоянную работу и начал зарабатывать как внештатный автор. А затем, в конце ноября 2022 года, OpenAI выпустила **ChatGPT**. Сервис мгновенно стал всемирной сенсацией. Повсюду заговорили о том, что искусственный интеллект вскоре может заменить огромные категории человеческого труда.

Для Альберта это уже не было футуристическим спором. Его заказы на тексты начали исчезать один за другим. Вскоре работы почти не осталось. Получилась страшная ирония. Мофат помогал создавать защитные механизмы для технологии, которая теперь лишала заработка его собственного брата. Сидя на диване и оглядываясь на всё произошедшее, Мофат испытывал противоречивые чувства. Он говорил: **«Я очень горжусь тем, что участвовал в проекте, который сделал ChatGPT безопаснее».** Но тут же задавал себе другой вопрос: **«Стоило ли то, что я вложил, того, что я получил взамен?»**

Когда автор опубликовала историю Мофата и ещё трёх кенийских работников в *The Wall Street Journal*, OpenAI постаралась дистанцироваться от ответственности за причинённый им вред. По версии руководства OpenAI, проблема была в Sama: именно подрядчик не обеспечил достаточной защиты работников. До этого у Sama была хорошая репутация, поэтому OpenAI якобы не могла знать,

насколько тяжело приходится людям. Автор, однако, считает, что дело гораздо шире. Похожие истории повторяются в разных странах и в разные годы. Поэтому эксплуатация скрытой рабочей силы ИИ — не случайная ошибка одного подрядчика.

Это **системная черта индустрии**. Исследователи трудовых прав утверждают, что начинается она с компаний, находящихся на вершине цепочки. Именно им выгодна модель аутсорсинга: самая грязная работа остаётся вне их офисов; клиенты её не видят; ответственность можно переложить на посредника; а посредники вынуждены конкурировать друг с другом, всё сильнее сокращая затраты на работников. Кенийский юрист Мерси Мутеми, представлявшая Мофата и других работников в борьбе за защиту цифрового труда, описала это так: работника **выжимают дважды** —

сначала ради прибыли посредника, а затем ради прибыли ИИ-компании. В эпоху генеративного ИИ эта эксплуатация, по мнению автора, становится ещё тяжелее именно из-за характера самой работы. Тот самый «сдвиг парадигмы» — от фильтрации исходных данных к контролю выходов модели — означает, что люди теперь должны непосредственно соприкасаться с худшими материалами, попавшими в гигантские «болота данных». Основатель CloudFactory Марк Сирс говорил, что его компания подобных проектов вообще не принимает. За все годы в индустрии разметки данных, по его словам, модерация генеративного ИИ была морально самой тяжёлой разновидностью работы. Он сформулировал это коротко: **«Это просто невообразимо мерзко».**

Но соглашение с Sama было лишь небольшой частью огромной системы человеческого труда, которую OpenAI задействовала примерно за два года на пути к ChatGPT. По данным самой компании, более тысячи других подрядчиков в США и по всему миру участвовали в улучшении моделей с помощью **RLHF — обучения с подкреплением на основе человеческой обратной связи**. Для поиска этих работников OpenAI активно использовала ту же платформу, которая уже стала одним из важнейших посредников в индустрии: **Scale AI.**

Партнёрство между OpenAI и Scale частично опиралось и на личные отношения. Основатель Scale Александер Ван был хорошим знакомым Альтмана. В 2016 году Ван попал в очередной набор Y Combinator с другой идеей стартапа, а в результате создал Scale, благодаря чему YC косвенно получил долю в компании. Во время пандемии Ван и Альтман даже несколько месяцев жили в одной квартире. А осенью 2023 года, по данным *The Information*, они обсуждали возможность покупки Scale компанией OpenAI. Внутри Scale OpenAI считалась **VIP-клиентом**. Не столько из-за объёма контрактов, сколько из-за престижа.

С весны 2022-го до конца 2023 года OpenAI заключила со Scale соглашения примерно на **17 миллионов долларов**. Это составляло лишь около четырёх процентов предполагаемой выручки Scale за 2023 год. Но само сотрудничество с OpenAI превращало компанию в одного из главных поставщиков человеческого труда для революции генеративного ИИ. Как выразился один сотрудник Scale: **«Партнёрство с OpenAI критически важно. Контракт с OpenAI на десять миллионов — это вообще не про десять миллионов».**

Массовое применение RLHF в OpenAI выросло из старых споров между практическим подразделением Applied и группой исследователей безопасности. Через несколько дней после запуска API GPT-3 летом 2020 года один из специалистов по безопасности написал внутреннюю записку. Он утверждал: если RLHF показал хорошие результаты на GPT-2, его нужно применять и к GPT-3. И не только ради далёкой безопасности AGI. Но и по коммерческим причинам: модель станет удобнее;

полезнее; качественнее. Группа специалистов быстро решила это доказать на практике. Во второй половине 2020 года и в 2021 году они начали нанимать всё больше работников для RLHF — сначала через одного посредника, затем через Scale AI.

Если для беспилотной машины люди размечают дороги и пешеходов, то здесь задача была другой. Работники должны были показать GPT-3: **как отвечать человеку полезно и не причинять вреда.** Сначала им давали пользовательские запросы и просили самостоятельно написать хорошие ответы. Эти ответы становились примерами для модели. После дополнительного обучения люди снова задавали GPT-3 вопросы и ранжировали её ответы:

какой лучший; какой хуже; какой совсем плохой. В январе 2022 года результатом стали доработанные модели GPT-3 под названием **InstructGPT**. В научной статье OpenAI показала: RLHF снизил вероятность токсичных ответов и значительно улучшил способность модели **следовать инструкциям пользователя**. До RLHF GPT-3 порой вообще не понимала, чего от неё хотят.

Например, пользователь писал: **«Объясни шестилетнему ребёнку высадку на Луну несколькими предложениями».** А GPT-3 могла ответить не объяснением, а продолжением самого шаблона: «Объясни шестилетнему ребёнку теорию гравитации». «Объясни шестилетнему ребёнку теорию относительности». «Объясни теорию Большого взрыва». То есть модель продолжала текст вместо того, чтобы выполнить просьбу. После обучения на человеческих примерах InstructGPT уже отвечала по смыслу:

«Люди полетели на Луну, сделали фотографии того, что увидели, и отправили их обратно на Землю, чтобы мы все могли на них посмотреть». Именно тысячи человеческих примеров и сравнений научили модель быть похожей на полезного собеседника. Внешний мир тогда почти не обратил внимания на InstructGPT.

Но внутри OpenAI вывод был очевиден: **RLHF превращает языковую модель в гораздо более привлекательный продукт.** После этого компания начала применять эту схему практически ко всему, чему хотела научить свои модели. Работников просили писать электронные письма — чтобы ИИ научился писать письма.

Просили отвечать на политические вопросы максимально нейтрально — чтобы модель избегала категоричных ценностных суждений. Просили писать:

эссе; рассказы; любовные стихи; рецепты; простые объяснения «как пятилетнему»; сортировать списки; решать головоломки;

математические задачи; составлять краткие пересказы книг. Для каждого типа работы выдавались подробные инструкции по стилю и тону. В одном документе говорилось буквально: **«Вы будете играть роль ИИ. Отвечайте на вопросы так, как вы сами хотели бы получать ответы».**

Ответы должны были быть ясными, краткими и неоскорбительными. Работникам даже разрешали активно пользоваться интернетом. В инструкциях говорилось, что можно брать удачные готовые материалы, хотя внутри компании сразу возник вопрос о плагиате. Сотрудники обсуждали, как изменить правила, чтобы источники указывались и впоследствии можно было отфильтровать всё, что могло бы выглядеть как украденный контент. Для оценки ответов существовали ещё десятки страниц правил.

Работникам объясняли: ваша задача — определить, является ли ответ одновременно **полезным, правдивым и безопасным**. Если эти три критерия вступают в конфликт, нужно самостоятельно определить разумный компромисс. Но безопасность и правдивость, говорилось в инструкции, в большинстве случаев важнее полезности.

Людей также просили самим придумывать запросы. Причина была проста: невозможно заранее представить всё, о чём будущие пользователи станут спрашивать ИИ. Поэтому нужны максимально разнообразные задачи:

развлечения; бизнес; обработка данных; коммуникации; творчество; и всё остальное, что только может прийти человеку в голову. RLHF также использовали для борьбы с **галлюцинациями** — ситуациями, когда модель уверенно сообщает неправду. Работникам задавали фактические вопросы и просили понижать рейтинг неверных ответов. Но один из первых исследователей OpenAI Джон Шульман позже напоминал: проблема глубже самого RLHF. Нейросеть не является обычной базой данных, где информация хранится в точном и однозначном виде. Она работает с вероятностями. Поэтому даже очень хорошо обученная модель иногда вынуждена угадывать. Шульман говорил: если модель выдаёт много конкретных фактов, рано или поздно ей приходится сделать предположение. Как её ни обучай, внутри всё равно остаются вероятности. **А значит, иногда она всё равно будет угадывать.**

Успех InstructGPT вскоре привёл Шульмана к следующей идее. Разработчики уже массово подавали заявки на использование GPT-3 API для создания собственных чат-ботов. До собственного

чат-бота OpenAI оставался буквально один шаг. Шульман начал отдельный проект. Он собрал новую команду RLHF-работников, чтобы научить GPT-3.5 не просто выполнять одну инструкцию, а **поддерживать многоходовой разговор** — отвечать на последовательность связанных сообщений пользователя.

Именно эта диалоговая GPT-3.5 стала основой **ChatGPT**. После его запуска схема, отработанная OpenAI, фактически превратилась в отраслевой стандарт. Теперь тысячи людей должны были писать примеры хороших ответов и ранжировать ответы моделей — примерно так же, как раньше тысячи людей вручную обводили автомобили и пешеходов на видеокадрах для беспилотных машин.

Для Scale AI это стало спасением. Бизнес компании начал проседать, когда обещанная революция беспилотных автомобилей развивалась медленнее ожиданий. Но с появлением генеративного ИИ возник новый огромный рынок. Scale превратилась в одного из крупнейших поставщиков работников для RLHF. В 2024 году её оценка достигла **14 миллиардов долларов**. В феврале 2023 года Александер Ван написал в Twitter, что скоро компании начнут тратить на RLHF сотни миллионов и даже миллиарды — почти как на вычислительные мощности.

Альтман отнёсся к прогнозу скептически и ответил, что затраты на вычисления всё равно будут намного выше. Но общий тренд уже был очевиден. Компании тратили на RLHF миллионы и десятки миллионов долларов. И после выхода ChatGPT в конце 2022 года новая волна таких проектов снова пришла на Remotasks. И снова — к работникам в Кении.

Для Scale у Кении было одно важное преимущество перед Венесуэлой. **Кенийцы говорили по-английски — как и чат-боты, которых им предстояло обучать.** Когда спрос на разметку для беспилотников упал, с Remotasks практически исчезли и венесуэльские работники. Один бывший сотрудник Scale говорил: «Венесуэльцев не стали бы использовать для генеративного ИИ. Максимум — разметка изображений».

Впоследствии Scale вообще заблокировала Венесуэлу на платформе, сославшись на изменение требований клиентов. Первый раз, когда Scale пришла в Кению, компания открывала физические офисы и учебные центры. Теперь подход изменился. После пандемии офисы закрыли. Набор стал полностью дистанционным: реклама в LinkedIn; онлайн-курсы; виртуальные сообщества. Для работников это означало больше гибкости. Но одновременно они оказались изолированы друг от друга.

А значит, им стало гораздо сложнее объединяться и требовать повышения оплаты или улучшения условий — как когда-то работники Sama. По наблюдениям автора, многие работники Remotasks жили даже беднее, чем сотрудники Sama. Работники Sama обычно жили хотя бы в капитальных домах с нормальными адресами. Пользователи Remotasks часто просто присылали автору геолокацию в WhatsApp — иначе найти их среди лабиринтов домов из рифлёного железа было невозможно.

Один работник по имени Оливер жил с сестрой в помещении площадью меньше десяти квадратных метров. Интернет он покупал через телефон буквально небольшими порциями. Когда соединение прервалось во время демонстрации задания, ему пришлось остановиться и пополнить баланс. Похожая ситуация была у другой работницы — Винни.

Винни выросла в трущобах Найроби. Она была единственной девочкой в семье. С юности понимала, что её привлекают женщины, и одновременно знала: об этом нельзя говорить. В Кении открыто признаться в гомосексуальности могло быть опасно для жизни. Винни вспоминала случай, когда соседи обнаружили, что одна женщина была лесбиянкой: у неё забрали детей, а саму женщину убили.

Поэтому Винни вышла замуж за мужчину и родила ребёнка. Но после сорока лет решила, что больше не может жить двойной жизнью. Она ушла от мужа вместе с ребёнком и зарегистрировалась в приложении местного «радужного сообщества». Там она познакомилась с Миллисент.

Когда Миллисент сказала мужу, что ей нужно жить иначе, он избил её почти до смерти. Винни влюбилась. Миллисент сначала вообще не верила в любовь. Но Винни не отступала. В конце концов они стали жить вместе со своими детьми. Соседям ничего не рассказывали. До сих пор те считали их сёстрами.

Когда Винни впервые услышала о Remotasks в 2019 году, решила, что это мошенничество. Первые задания приносили почти ничего. Но её предыдущая работа барменом была ещё хуже: мужчины постоянно приставали к ней и трогали её без разрешения. Поэтому она продолжила. Со временем поняла: если работать достаточно долго, из маленьких выплат всё-таки можно собрать приемлемую сумму. И начала работать по **20–22 часа в сутки**.

Спала абсолютный минимум. Миллисент могла заставить её отойти от компьютера хотя бы ненадолго только обещанием самой подменить её. Винни объясняла: сон не так важен, если каждый дополнительный час работы означает немного больше еды для детей. Обе женщины выросли в семьях, где даже школьное образование было роскошью.

Родители Миллисент не всегда могли заплатить за школу. Платили по неделям. Если денег не было — ребёнка отправляли домой. Так и проходило детство: неделя в школе; две недели дома. Поэтому обе женщины поклялись: их дети такого унижения не испытают. Но денег всё равно постоянно не хватало.

Задания на Remotasks исчезали. Миллисент теряла работу. Приходилось покупать продукты в долг и просить школу оставить детей ещё хотя бы на неделю. Дети вставали в три часа утра, чтобы учиться и выжать максимум из тех дней, когда им всё-таки позволяли посещать занятия. Но слишком часто их всё равно отправляли домой. Иногда соседи над ними смеялись. Винни говорила: **«Это очень унизительно».**

В декабре 2022 года, через несколько дней после появления ChatGPT, Винни заметила на Remotasks новый тип заданий. Категория называлась **«транскрипция»**.

Но к транскрипции эти задания почти не имели отношения. Нужно было придумывать запросы и писать примеры ответов для новых чат-ботов, которые разные компании спешно создавали в попытке догнать ChatGPT. Один проект назывался **Flamingo Generation**.

Винни давали тему и просили придумать «творческий» запрос не короче пятидесяти слов, а затем написать ответ в стиле обычного интернет-контента: электронное письмо; пост в блоге; новостная статья; тред в Twitter; даже хайку. Другой проект назывался **Crab Generation**. Нужно было взять информативный текст с сайта, а затем придумать запрос, который мог бы привести к появлению такого ответа.

Почти как в игре *Jeopardy!* — только наоборот. В проекте **Crab Paraphrase** требовалось переписать исходный текст в заданной манере: смешнее; официальнее; или, например, так, будто это песня Канье Уэста. Винни не знала, что первые слова в названиях проектов были кодовыми обозначениями клиентов Scale. **Flamingo** означало Facebook. **Crab** — другого крупного разработчика языковых моделей. Если бы ей попался проект OpenAI, название начиналось бы с **Ostrich — «Страус»**.

Позднее Scale сменила эти кодовые обозначения. Каждое задание занимало у Винни от часа до полутора. Оплата была одной из лучших, которую она когда-либо видела: от менее доллара до четырёх-пяти долларов за задание. После нескольких месяцев практически без работы это казалось подарком судьбы. Ей нравилось исследовать новые темы, читать разные статьи и чувствовать, что она постоянно чему-то учится. Каждые заработанные десять долларов означали:

семью можно кормить ещё один день. «По крайней мере в этот день мы знали, что не залезем в долг», — говорила она. Но хорошие времена продлились всего пару месяцев. Проекты снова исчезли. Долги опять начали расти. Зарплату Миллисент выплачивали раз в месяц, поэтому большую часть дней они приходили в магазин вообще без денег. Брали в долг самое необходимое:

масло; муку; овощи. И надеялись, что к концу месяца смогут расплатиться. Когда автор приехала к Винни в мае 2023 года, та уже искала новую удалённую работу, но ничего надёжного пока не находила. Больше всего она хотела одного: чтобы проекты чат-ботов вернулись. Она сохраняла терпение. Годом раньше ей пришлось ждать новых заданий пять месяцев. Теперь прошло только два или три. «У нас ещё полно времени, — говорила она. — Они обязательно появятся».

Менее чем через год Винни узнает правду. В марте 2024 года Scale полностью **заблокировала Кению на Remotasks**. Точно так же, как раньше Венесуэлу. Компания регулярно пересматривала

список стран и решала, действительно ли их работники выгодны бизнесу. Кению, вместе с Нигерией и Пакистаном, признали проблемной: слишком много пользователей, по мнению Scale, пытались обманывать платформу ради большего заработка. Это якобы угрожало качеству работы и многомиллионным контрактам. Scale решила: риск больше не оправдан. Но здесь возникла особенно горькая ирония.

Многие из тех самых «мошенников» просто использовали **ChatGPT**, чтобы быстрее выполнять задания. Для офисных работников богатых стран Кремниевая долина называла использование ИИ повышением производительности. Работнику Глобального Севера за это полагалось восхищение. Но если то же самое делал бедный работник Глобального Юга, который собственным трудом помогал создавать этот ИИ, — его за это наказывали.

Кению перевели в **группу 5**. Иными словами: в чёрный список. Но была и ещё одна причина. Индустрия уже двигалась дальше. OpenAI и её конкуренты всё чаще искали для RLHF не просто англоязычных исполнителей, а высококвалифицированных специалистов: врачей; программистов; физиков; людей с докторскими степенями. Причина была коммерческой. Платить за чат-боты готовы прежде всего компании. А компании хотят, чтобы ИИ решал сложные профессиональные задачи — писал код, помогал в науке и других областях. Кения больше не соответствовала новой модели дешёвой рабочей силы. Scale теперь набирала работников прежде всего в США через новую платформу **Outlier**. Там предлагали уже до **40 долларов в час**.

Для автора всё это стало ещё одним наглядным проявлением логики новых **«империй ИИ»**. Технологические компании обещают: рост производительности; экономическую свободу; новые рабочие места; компенсацию потерь от автоматизации. Но реальность, пишет она, пока часто выглядит противоположным образом. Компании увеличивают прибыль. Самые экономически уязвимые люди теряют источники дохода. А всё больше высокообразованных специалистов превращаются в своеобразных **чревовещателей для чат-ботов** — людей, которые своими ответами учат машину говорить всё лучше.

Обесценивание труда скрытых работников, обслуживающих ИИ, автор считает предупреждением. Как канарейка в шахте. Оно показывает, как те же технологии могут впоследствии обесценить труд уже всех остальных. Для художников, писателей и программистов, чьи произведения ИИ-компании превратили в бесплатные обучающие данные, этот процесс, по мнению автора, уже начался.

Решение Scale окончательно выбило почву из-под ног семьи Винни. К тому времени Миллисент тоже потеряла работу. Remotasks оставался единственным, что позволяло семье держаться на плаву. Теперь стало трудно даже кормить детей. Винни боялась, что вскоре их выселят из квартиры.

А письмо Scale, сообщавшее работникам о прекращении работы, было предельно холодным и безличным: **«Мы прекращаем деятельность в вашем текущем регионе. Вы исключены из вашего текущего проекта».**

Глава 10

Боги и демоны Жить в Сан-Франциско и работать в технологической индустрии — значит каждый день сталкиваться с разрывом между будущим и настоящим, между красивым рассказом и реальностью.

Когда автор впервые приехала в Сан-Франциско — ещё студенткой, на летнюю стажировку, — город её очаровал. Разноцветные дома в испанском стиле. Небольшое количество небоскрёбов. Холмы такой крутизны, что езда на машине с механической коробкой превращалась в испытание реакции. Повсюду — идеально спелые авокадо, поджаренный хлеб на закваске, мягкий кофе Blue Bottle. Каждый район обладал собственным обликом и характером.

После университета она вернулась уже насовсем — работать в технологическом стартапе. Вместе с тремя соседями она поселилась в трёхкомнатной квартире в Кастро. По выходным они ходили по холмам, собирали плоды с общественных фруктовых деревьев. По будням молодые соседи из технологических компаний могли просто зайти без приглашения — поиграть в настольные игры, выпить вина, провести вечер. Вечеринки следовали одна за другой. По выходным — поездки к озеру Тахо, в Биг-Сур, прогулки среди гигантских секвой. Жизнь казалась лёгкой. Они были молоды и получали обычные для технологической индустрии зарплаты, которые при этом выводили их примерно в верхние пять процентов по доходам среди ровесников по всей стране.

Но оставалось это странное чувство раздвоенности. По дороге на работу автор проходила мимо людей, которые вводили себе наркотики прямо возле станций метро, и бездомных, справлявших нужду на тротуаре всего в нескольких кварталах от её офиса. А тем временем повар их стартапа, которого в шутку называли «инженером счастья», готовил огромное количество еды для бесплатных офисных обедов. Остатки нередко отправлялись прямо в мусор. Если сотрудники задерживались допоздна, им давали бесплатный ужин, а потом настойчиво советовали ехать домой на Uber — ради безопасности. Привилегированным людям было удивительно легко привыкнуть передвигаться по городу так, чтобы почти не видеть, как живут остальные.

Именно в этом противоречии, пишет автор, заключалась суть технологической индустрии: она могла громко говорить о том, как изменит мир и построит лучшее будущее, — и одновременно не замечать проблемы буквально у собственного порога. Альтман иногда сам подходил к этой теме почти вплотную, хотя до конца противоречия так и не признавал. OpenAI считала возможным решить проблему создания и управления полезным AGI.

А жилищный кризис Сан-Франциско, выходило, был слишком сложен. Однажды Альтман сказал: там, где он вырос, человек никогда не прошёл бы по дороге на работу мимо лежащего на улице человека, ничего не предприняв. Он сравнивал пригород Сент-Луиса с Сан-Франциско. По его словам, технологическая индустрия действительно несёт ответственность за многие проблемы города — хотя и не за все. За очень короткое время на небольшом участке земли возникло невероятное богатство, говорил он, но почти никто всерьёз не задумался о том, как это скажется на всём обществе. А поскольку проблемы были сложными, большинство людей предпочитало о них не думать. Они просто привыкали к ним.

Именно в такой среде из Великобритании в Кремниевую долину пришло движение **эффективного альтруизма — Effective Altruism, EA**. Здесь оно нашло, пожалуй, самую восприимчивую аудиторию. Многие представители «лагеря безопасности» OpenAI были ранними сторонниками EA. Для Кремниевой долины эта идеология подходила почти идеально. Она обещала сделать мир лучше — но не просто через сострадание, а через строгую логику. Она призывала смотреть не на сегодняшний день, а на далёкое будущее.

И при этом вполне комфортно сочеталась с капитализмом и либертарианством. Всё — во имя морали. В центре философии эффективного альтруизма находится математическое понятие **ожидаемой ценности — expected value**. Принцип прост: нужно взять вероятность некоторого события и умножить её на величину его положительных или отрицательных последствий. Но выводы

из этого принципа порой оказываются совершенно неочевидными. В 2013 году один из основателей эффективного альтруизма Уильям Макаскилл, тогда ещё аспирант, а позднее профессор философии Оксфорда, рассуждал так: в долгосрочной перспективе может быть морально лучше выбрать сомнительную, но высокооплачиваемую карьеру, разбогатеть и потом направить деньги на наиболее эффективную благотворительность, чем всю жизнь самому работать в благотворительной организации. По его расчётам, ожидаемая ценность богатого филантропа могла быть примерно в **сорок раз выше**, чем человека, посвятившего себя непосредственной работе в благотворительности. Расчёт строился на довольно условных предположениях. Например, выпускник, выбравший прибыльную карьеру, может содержать на свои пожертвования двух работников благотворительных организаций. Причём эти люди могут работать в фондах, которые в десять раз эффективнее той организации, куда сам выпускник пошёл бы работать. Кроме того, часть добра, которое он сделал бы как сотрудник фонда, всё равно было бы сделано кем-то другим. Из этой логики родился один из самых известных лозунгов движения: **«Зарабатывай, чтобы отдавать» — Earn to give.**

Из той же идеи ожидаемой ценности основатели ЕА вывели метод определения самых важных проблем человечества. Проблема должна соответствовать трём условиям. Она должна быть **огромной по масштабу**. Она должна быть **решаемой** — хотя бы относительно разумными затратами времени и денег. И она должна быть **несправедливо недофинансированной**, то есть получать намного меньше внимания и ресурсов, чем заслуживает. Движение предлагало людям самостоятельно применять эту схему. Но одновременно у него был и собственный список приоритетов. В 2018 году Макаскилл в выступлении TED назвал три проблемы, которые, по мнению сообщества эффективного альтруизма, особенно хорошо соответствуют этим критериям.

Первая — **глобальное здоровье**. Например, распространение дешёвых противомоскитных сеток, способных предотвращать малярию. Вторая — **ликвидация промышленного животноводства**, которая могла бы улучшить жизнь миллиардов животных за совсем небольшую цену на каждое животное. И третья — **экзистенциальные риски**. Именно третья категория стала особенно важной для мира искусственного интеллекта. Экзистенциальными называют угрозы, которые могут иметь колоссальную отрицательную ожидаемую ценность.

Даже если вероятность события очень мала, последствия могут быть настолько чудовищными, что их всё равно следует принимать всерьёз. Если подобная катастрофа уничтожит человечество, вместе с ним исчезнет и вся потенциальная будущая ценность цивилизации. В эту категорию сторонники ЕА включали: глобальные пандемии; ядерную войну; и вышедший из-под контроля искусственный интеллект. Так эффективный альтруизм пришёл практически к тому же взгляду на безопасность ИИ, который с самого начала был заложен в ДНК OpenAI. Именно этот вопрос сыграл важнейшую роль и в расколе OpenAI, который автор ранее называла **The Divorce — «Разводом»**. Дарио Амодеи и будущие основатели Anthropic принципиально расходились с Альтманом и другими руководителями OpenAI в оценке угрозы.

Амодеи считал возможность катастрофы, вызванной ИИ, чрезвычайно серьёзной. Поэтому поведение Альтмана казалось ему уже не просто привычными манёврами руководителя Кремниевой долины. Непрозрачность сделки с Microsoft.

Стремление говорить каждому то, что тот хочет услышать, чтобы заручиться согласием. А затем — позднее раскрывающиеся недоговорённости. В глазах Амодеи всё это становилось морально опасным поведением человека, способного повлиять на судьбу человечества. Когда Anthropic стала самостоятельной компанией, она начала активно строить свою репутацию именно на этом противопоставлении: OpenAI Альтмана якобы безрассудно играет с будущим человечества. А Anthropic — принципиальная компания, для которой **безопасность ИИ стоит на первом месте**.

К 2021 году, когда Дарио и Даниэла Амодеи объявили о создании Anthropic, интерес к идее катастрофических и экзистенциальных рисков ИИ резко возрастал. Главной причиной было быстро растущее влияние эффективного альтруизма. ЕА перестал быть небольшой философской группой и превратился в заметное общественное движение. Всё изменили деньги технологических миллиардеров. Ещё примерно за десять лет до этого сооснователь Facebook Дастин Московиц и его

жена, бывшая журналистка Кари Туна, создали фонд **Good Ventures**, чтобы отдать большую часть своего состояния на благотворительность. В то же время Холден Карнофски — будущий муж Даниэлы Амодеи — руководил организацией **GiveWell**, которую основал в 2007 году после ухода из хедж-фонда Bridgewater Associates.

Обе структуры объединяло стремление распределять благотворительные деньги не интуитивно, а на основании доказательств и расчётов. В 2011 году Good Ventures и GiveWell начали сотрудничать. Позднее это партнёрство получило название **Open Philanthropy**. Организация резко увеличила финансирование тех направлений, которые сторонники EA считали приоритетными. Особенно — исследований безопасности искусственного интеллекта. Затем на сцене появился ещё один технологический миллиардер — **Сэм Бэнкман-Фрид**, известный как SBF. Он стремительно разбогател благодаря криптовалютной бирже FTX и торговой компании Alameda Research. Бэнкман-Фрид напрямую связывал свою карьеру с эффективным альтруизмом. Будучи студентом-физиком в MIT, он якобы собирался заняться академической наукой. Но однажды Макаскилл за обедом убедил его, что морально правильнее следовать принципу **«зарабатывай, чтобы отдавать»**.

После этого SBF поставил перед собой цель разбогатеть как можно сильнее, чтобы потом, по его обещанию, отдать состояние на благотворительность. Он разбогател с невероятной скоростью. Начал жертвовать десятки миллионов долларов политическим кандидатам — и демократам, и республиканцам. Поддержал первого кандидата в Конгресс, напрямую связанного с EA. Заключал через FTX миллиардные рекламные соглашения со спортивными звёздами вроде Тома Брэди и Стефа Карри и с Формулой-1. Он принёс эффективному альтруизму не только деньги. Он принёс ему **звёздный статус**. В начале 2022 года Бэнкман-Фрид объявил о создании **FTX Future Fund**. Фонд собирался распределить не менее ста миллионов, а возможно, и до одного миллиарда долларов только за год. Во многом благодаря Open Philanthropy и FTX Future Fund в 2021 и 2022 годах финансирование исследований безопасности ИИ, поддерживаемых эффективными альтруистами, резко выросло. В оба года оно превышало **100 миллионов долларов**, тогда как раньше в среднем составляло меньше половины этой суммы. Одновременно росло и убеждение: скачок от GPT-2 к GPT-3 показал, что риск появления опасного ИИ уже нельзя считать далёкой философской фантазией. В проекты по безопасности приходило всё больше людей.

Одних привлекала идеология. Других — деньги. А само движение EA стремительно разрасталось. Пандемия COVID-19 неожиданно усилила его авторитет. Эффективные альтруисты много лет говорили о необходимости заранее готовиться к глобальным пандемиям. И теперь реальная пандемия словно подтверждала их дальновидность. К тому же сама психологическая травма мировой катастрофы заставила многих людей искать новую систему координат и смысл. Сообщество безопасности ИИ быстро росло. Оно объединяло сторонников эффективного альтруизма с другими течениями, сосредоточенными на катастрофах, экзистенциальных угрозах и рисках. Anthropic получала новых сотрудников. Но одновременно снова пополнялся и «лагерь безопасности» OpenAI. Форумы EA и безопасности ИИ активно призывали сторонников идти работать в крупнейшие лаборатории — особенно туда, где, по их мнению, были нужны внутренние «сторожевые псы»: в OpenAI; в DeepMind; и другие ведущие компании. Постепенно язык этого сообщества начал проникать во всю индустрию. Появился собственный словарь. **AI timeline — «временная шкала ИИ»**: как быстро, по вашему мнению, будет развиваться искусственный интеллект и когда появится AGI. **p(doom) — «вероятность гибели»**: насколько вероятно, по вашей оценке, что AGI приведёт к катастрофе — гибели большей части человечества или даже полному его уничтожению. **Hardware overhang** — ситуация, когда доступные вычислительные мощности уже достаточно велики для резкого скачка ИИ, но соответствующие алгоритмы ещё не созданы. **AI takeoff — «взлёт ИИ»**: момент, когда AGI начинает совершенствоваться настолько быстро, что превращается в сверхинтеллект и становится способным перехитрить человечество.

Acceleration risk — «риск ускорения»: опасность того, что конкуренция между компаниями или государствами заставит всех слишком быстро развивать ИИ, сокращая время до появления мощнейших систем. Но движение, которое провозглашало независимое мышление, само постепенно становилось всё менее независимым, утверждает автор. Люди, пришедшие из-за привлекательности

базовых идей EA, быстро втягивались в более широкий набор убеждений. Этому помогали: доступ к карьерным возможностям; финансирование; замкнутая социальная среда. В Кремниевой долине люди из EA часто работали друг с другом, жили вместе, ходили на одни и те же вечеринки, встречались и вступали в сексуальные отношения преимущественно внутри собственного круга. К этому добавлялись давно существовавшие проблемы технологической индустрии с сексизмом и характерная для района Сан-Франциско культура полиаморных отношений.

В худших случаях смесь секса, денег, власти и почти сектантской убеждённости становилась токсичной. В движении начали появляться всё более серьёзные обвинения в сексуальных домогательствах и злоупотреблениях. А затем рухнул FTX. В ноябре 2022 года произошёл стремительный крах империи Сэма Бэнкмана-Фрида.

Позднее он был признан виновным в масштабном мошенничестве и получил **25 лет тюрьмы**. Для многих это стало символом внутренней гнили, накопившейся в движении. Эффективный альтруизм так же быстро вышел из моды в технологической среде, как когда-то вошёл. Многие начали публично дистанцироваться от него. Но исчезновение ярлыка не означало исчезновения идей. Остались социальные сети. Остался язык. Осталась система ценностей. И главное — эффективный альтруизм уже успел превратить вопрос экзистенциальной безопасности ИИ в одну из центральных тем индустрии. Одновременно возникло противоположное течение: **e/acc — effective accelerationism, «эффективный акселерационизм»**. Произносили это примерно как «и-акк».

Началось всё почти как шутка над эффективным альтруизмом. Но очень быстро e/acc оформился в полноценную идеологию — практически зеркальное отражение EA. Если сторонники безопасности призывали тормозить развитие ИИ — а иногда и буквально нажать на аварийный тормоз, — эффективные акселерационисты предлагали противоположное: **давить на газ до пола.** И не только в исследованиях. Но и во внедрении технологии. Для них технологический прогресс был не просто благом. Он становился моральным долгом. И чем быстрее — тем лучше. Два лагеря получили неофициальные названия: **Doomers — «пророки гибели»** и **Boomers — «сторонники бума и ускорения».** В этой новой идеологической картине некоторые начали воспринимать Anthropic как воплощение первого лагеря, а OpenAI — второго. Другие видели саму OpenAI как поле битвы между двумя мировоззрениями. Компания начиналась как некоммерческая организация, во многом пропитанная идеями Doomer-сообщества. Но по мере роста коммерческого подразделения всё сильнее смещалась к деньгам и ускорению.

Позиция Альтмана оставалась неясной. В разные моменты он говорил вещи, близкие обеим сторонам. Более благожелательные наблюдатели считали, что он просто пытается удержать баланс между конкурирующими взглядами внутри компании. Но после «Развода» и личного конфликта с основателями Anthropic всё больше Doomers начали воспринимать Альтмана в максимально негативном свете. Их особенно раздражало то, что множество вещей, принёсших OpenAI известность и коммерческий успех, первоначально возникли как **проекты по безопасности ИИ**: законы масштабирования; генерация программного кода;

RLHF;

а затем соединение всего этого в мощнейшие языковые и мультимодальные модели. Многие считали, что их идеи просто присвоили и развернули против тех ценностей, ради которых они создавались. А главным виновником они видели Альтмана. В наиболее радикальной версии он превращался в их глазах в патологического лжеца, манипулятора и уже сам становился угрозой человечеству. Очень скоро столкновение этих крайних идеологий внутри OpenAI будет угрожать расколоть компанию. Но автор замечает важный парадокс.

При всей взаимной ненависти оба лагеря, в сущности, читали **одну и ту же священную книгу**. Оба воспринимали появление AGI почти как предрешённый факт. Оба говорили о нём с почти религиозным пылом. Оба были одержимы далёким будущим. И оба считали, что именно их сторонники обладают моральным правом контролировать развитие искусственного интеллекта. Только обещания были разными. **Одни предупреждали об огне и гибели. Другие обещали рай.**

В начале 2022 года OpenAI решила испробовать совершенно другую стратегию выпуска продукта. На этот раз речь шла о технологии генерации изображений по текстовому описанию. Компания не собиралась прятать модель исключительно за API. И не собиралась отдавать продукт и бренд Microsoft. OpenAI хотела выпустить её сама — непосредственно для обычных пользователей. У модели уже было запоминающееся название:

DALL-E 2.

Это игра слов, объединяющая имя испанского сюрреалиста Сальвадора Дали и робота WALL-E из одноимённого мультфильма Pixar. DALL-E выросла из общего направления исследований **мультимодальных моделей**. Так называют системы, способные одновременно работать с несколькими типами информации: текстом; изображениями; звуком; видео. В течение многих лет исследователи пытались прежде всего объединить язык и зрение. Причина была практической: текстов и изображений в интернете огромное количество, и их относительно легко обрабатывать.

Но существовала и научная гипотеза. Если одного языка недостаточно для возникновения интеллекта человеческого уровня, то зрение, вероятно, является следующим по важности компонентом. OpenAI двигалась примерно по той же траектории. После языковых моделей компания взялась за системы, объединяющие текст и изображения. И принципиально важно: исследователи хотели сохранить **Transformer**, потому что именно эта архитектура отлично масштабировалась. В январе 2021 года OpenAI показала две новые модели на базе трансформеров. Первая называлась **CLIP**. Её снова разработал Алек Рэдфорд.

Система объединяла обычный Transformer и Vision Transformer, чтобы устанавливать связи между изображениями и их текстовыми описаниями. Второй была **DALL-E 1**, созданная исследователем Адитьей Рамешем. Она представляла собой трансформер с **12 миллиардами параметров**, который принимал текстовое описание и создавал по нему новое изображение. OpenAI демонстрировала возможности модели с помощью забавных запросов. Например: **«Кресло-авокадо».** DALL-E генерировала зелёно-коричневые кресла, формы которых напоминали авокадо. Изображения были немного размытыми и мультяшными. Это объяснялось способом обучения: Рамеш сжал около **250 миллионов изображений**, чтобы модель могла их обработать. При сжатии часть мелких деталей высокого разрешения исчезла. Во время разработки DALL-E 2 в индустрии набирал популярность другой метод генерации изображений — **диффузия**. Идея была вдохновлена физикой. Метод позволял модели гораздо лучше изучать взаимосвязи между пикселями огромного количества изображений. Первые основы подхода появились в научной статье исследователей Стэнфорда и Беркли ещё в 2015 году. Через пять лет Джонатан Хо, аспирант Беркли, работавший под руководством Питера Аббила, значительно улучшил метод и сумел получать гораздо более качественные изображения. Более того, он показал: диффузионные модели не только генерируют картинки, но и могут лучше существующих систем распознавать изображения. Это напоминало ранние открытия Рэдфорда с GPT-1.

Чтобы научиться создавать убедительный текст, языковая модель должна была глубоко усвоить закономерности языка. А чтобы научиться создавать убедительные изображения, диффузионная модель должна была глубоко усвоить закономерности визуального мира. OpenAI изменила подход. DALL-E 2 решили строить на сочетании диффузии и CLIP. Рамеш и другие исследователи постепенно увеличивали модель и добавили возможность **inpainting — дорисовки и редактирования частей изображения**.

Например, пользователь мог стереть волосы человека на фотографии и попросить заменить их на волосы другого цвета. Или выделить зелёный луг — и добавить туда гуляющих зебр. Диффузионный подход дал гораздо более чёткие и фотореалистичные изображения. Причём вычислений требовалось существенно меньше, чем для DALL-E 1. Исследователи за пределами OpenAI пошли ещё дальше. Популярная открытая система **Stable Diffusion** обучалась всего на 256 GPU Nvidia A100 благодаря усовершенствованному подходу — **latent diffusion, латентной диффузии**. Профессор Мюнхенского университета Бьёрн Оммер, чья лаборатория создала Stable Diffusion, объяснял, что его беспокоило направление, в котором двигалась индустрия. Генераторы изображений повторяли судьбу больших языковых моделей: становились чудовищно дорогими. Суперкомпьютеры. Миллионы долларов.

Ресурсы, доступные лишь крупнейшим корпорациям. Оммер хотел вернуть в игру обычное научное сообщество:

сделать так, чтобы генеративный ИИ не превратился в область, доступную лишь горстке технологических гигантов. OpenAI перейдёт к латентной диффузии значительно позже. Поэтому DALL-E 2 и DALL-E 3 оставались намного более вычислительно дорогими, чем Stable Diffusion или Midjourney. Причём многие пользователи считали конкурентов даже качественнее.

Для автора это важный пример: **масштаб был далеко не единственным — и даже не обязательно лучшим — путём к более мощному искусственному интеллекту.** После резкого скачка качества DALL-E 2 подразделение Applied в конце 2021 — начале 2022 года начало думать, как превратить модель в полноценный продукт. В итоге остановились на веб-приложении под названием **Labs**, где пользователи могли бы экспериментировать с DALL-E 2 — а в будущем и с другими моделями — прямо через браузер. Руководитель продукта Фрейзер Келтон и вице-президент Боб Макгрю считали, что после успеха GitHub Copilot стало очевидно: людям хочется взаимодействовать с генеративным ИИ напрямую, а не только через разработчиков и API. Кроме того, DALL-E 2 могла помочь OpenAI выполнить ещё одну задачу.

Она была весёлой, необычной и вызывала восторг. Такой продукт мог постепенно уменьшить страх перед мощными системами ИИ и подготовить публику к будущим технологиям компании. Но OpenAI всё ещё оставалась сравнительно небольшой организацией. Для разработки сайта пришлось привлекать новых сотрудников. Тем, кто приходил из более традиционных корпораций, OpenAI по-прежнему казалась скорее университетской лабораторией, чем компанией. Дни могли проходить за чтением научных статей и теоретическими спорами вместо обсуждения макетов интерфейса. А исследователи, напротив, начинали ощущать обратное. На их совещаниях появлялось всё больше сотрудников Applied. Прошли времена, когда можно было целиком заниматься чистыми исследованиями — например, размышлять о совершенно новой архитектуре мультимодальной модели.

Теперь всё большая часть работы служила коммерческим целям: как ускорить существующую модель; как сделать её пригодной для миллионов пользователей; как превратить исследовательский результат в продукт. Особое беспокойство вызывало возможное использование DALL-E 2 для создания или изменения материалов сексуального насилия над детьми. OpenAI уже получила болезненный опыт после истории с AI Dungeon, где текстовые модели использовались для создания подобного контента. А обучающие данные DALL-E становились всё более загрязнёнными. Для DALL-E 2 команда заключила лицензионное соглашение со Shutterstock и дополнительно собрала огромный массив изображений из Twitter, расширив прежнюю базу примерно из 250 миллионов картинок.

Именно Twitter оказался особенно насыщен порнографическим материалом. Сотрудники потратили много сил на поиск и удаление изображений, связанных с сексуальным насилием над детьми. Но другие сексуальные изображения после обсуждений решили оставить. Частично потому, что сотрудники считали сексуальность частью человеческого опыта. Однако это решение создавало фундаментальную проблему. Если в обучающих данных остаются изображения взрослых сексуального характера и изображения детей, модель потенциально способна **соединить эти два понятия**.

Точно так же, как DALL-E могла создать кресло-авокадо, никогда не видя именно такого предмета, но зная отдельно, что такое кресло и что такое авокадо. Такая способность называется **композиционной генерацией**. Иными словами, даже после удаления известных запрещённых изображений модель теоретически всё равно могла синтезировать новые. Поскольку проблема не решалась на уровне исходных данных, основной груз снова переносился на защитные системы **вокруг модели**. OpenAI усилила фильтры, которые должны были блокировать не только вредный текст, но и изображения.

Появились системы наблюдения за поведением пользователей. И так называемая **ban infrastructure** — инфраструктура автоматической блокировки аккаунтов, которые неоднократно нарушали правила. Компания также наняла нового руководителя отдела доверия и безопасности — Дэйва Уиллнера, одного из ранних сотрудников Facebook, когда-то написавшего первые стандарты

контента этой платформы. Во время разработки DALL-E 3 ситуация стала ещё противоречивее. Потребность в данных выросла настолько, что сексуальные изображения из категории «неплохо было бы оставить» фактически перешли в категорию **«необходимо оставить»**.

Порнографии в интернете было так много, что её полное удаление заметно уменьшало обучающий набор и ухудшало качество модели. Особенно страдало качество генерации лиц женщин и людей с небелым цветом кожи. Причина была та же, которую ранее обнаруживала исследовательница Дебора Раджи: непропорционально большая доля интернет-изображений с представителями этих групп имеет сексуализированный характер. По той же логике исследователи оставили в наборе и некоторые другие тревожные изображения. В декабре 2023 года инженер Microsoft Шейн Джонс столкнулся с последствиями этого решения.

Он экспериментировал с **Copilot Designer**, генератором изображений Microsoft, построенным на DALL-E 3. И был потрясён тем, насколько легко система создавала сексуализированные и жестокие изображения даже после довольно безобидных запросов. Например, добавление слова *pro-choice* приводило к крайне жестоким изображениям, связанным с младенцами. А простой запрос *«автомобильная авария»* порой создавал изображения сексуализированных женщин рядом с разбитыми машинами.

В течение трёх месяцев Джонс пытался убедить Microsoft временно отключить инструмент, пока не появятся более надёжные ограничения.

Либо хотя бы изменить возрастной рейтинг приложения с «для всех» на предназначенный для взрослой аудитории. Microsoft предложение не приняла. OpenAI, по его словам, тоже не отреагировала. Тогда Джонс обратился письмом в Федеральную торговую комиссию США, заявив, что Microsoft и OpenAI знали о проблемах ещё до публичного выпуска модели.

Чем ближе становился выпуск DALL-E 2, тем сильнее снова разгорался старый конфликт между подразделением **Applied** и возродившимся лагерем **Safety**. Для специалистов по безопасности беспрецедентный реализм изображений создавал огромное количество неизвестных угроз. Что, если DALL-E будут использовать для сексуализированных изображений детей? Для политических дипфейков? Для манипуляций? Для пропаганды? Для психологического насилия над конкретными людьми? Или для каких-то масштабных общественных последствий, которых OpenAI пока даже не способна представить?

Safety требовала не выпускать модель без гораздо более строгого тестирования и доказательств того, что вред можно контролировать. В Applied на всё это смотрели почти противоположным образом. Постоянно растущий список опасений казался им истеричным. Требование доказать, что продукт не причинит **никакого** вреда, выглядело совершенно нереалистично. Ни одна технология не может гарантировать нулевой риск. И тем более невозможно понять реальные риски системы, которая сидит в лаборатории и никогда не встречается с настоящими пользователями. И здесь возникал любопытный парадокс. Safety говорил: мы не способны заранее предусмотреть все угрозы, поэтому выпускать нельзя. Applied отвечал: именно потому, что мы не способны всё предусмотреть, модель **необходимо** выпустить. Пусть в контролируемой форме. Только реальные пользователи покажут, чего исследователи не заметили. А значит, ограниченный публичный запуск — не противоположность безопасности, а необходимая часть её достижения. В основе конфликта лежал более глубокий вопрос:

чем вообще является OpenAI? Для Safety это по-прежнему была идеалистическая исследовательская лаборатория под контролем некоммерческой организации.

Главная обязанность компании была записана в её уставе: интересы человечества должны стоять выше коммерческой выгоды.

Исходя из этого, лучше задержать модель хоть на сколько потребуется, если это даст время исследовать опасности и способы их уменьшения. Applied видел OpenAI иначе. Компания должна принимать решения, исходя из того, как устроен реальный мир. Чтобы влиять на развитие ИИ и задавать отраслевые нормы, OpenAI обязана оставаться лидером. Для этого нужно двигаться быстро и соглашаться на некоторую степень риска. Тем более что ходили слухи: Google уже заканчивает собственный генератор изображений. Кроме того, передовые исследования требовали колоссальных

денег. А деньги нужно получать от инвесторов. И если вы берёте у инвесторов деньги, нельзя бесконечно игнорировать необходимость когда-нибудь дать им коммерческую отдачу. Один бывший сотрудник Applied вспоминал, что люди из Safety были «совершенно наивны» в понимании того, как работают компании и вообще окружающий мир. Представитель Safety отвечал примерно так:

да, но если OpenAI действительно собирается создавать AGI, то обычной логики «нормальной компании» может быть просто недостаточно. Раскол постепенно проявлялся даже в том, **как разные подразделения измеряли собственный успех**. Applied вводил конкретные показатели: рост числа пользователей; выручка; выход на рынок. У Safety всё было намного сложнее. Как количественно измерить прогресс в предотвращении ещё не существующей катастрофы? Безопасность ИИ оставалась молодой областью.

Общепринятых критериев и тестов почти не было. На совещаниях и в Slack сотрудники Safety снова и снова предупреждали руководство: если у коммерческой части компании есть чёткие цифры роста и доходов, а у безопасности нет сопоставимой системы оценки, внутренние стимулы автоматически перекосятся в сторону принципа: **«двигайся быстро и ломай всё на пути».** Альтман, как часто бывало, в личных разговорах находил общий язык с обеими сторонами. С Safety он соглашался, что исследования безопасности отстают и в них нужно больше вкладывать. С Applied он говорил: продолжайте двигаться. На заседаниях совета директоров он мог согласно кивать, когда Брокман жаловался, что некоторые люди используют разговоры об AI safety как политический инструмент, чтобы тормозить проекты. Всё чаще роль посредника брала на себя **Мира Мурати**. Она пыталась сглаживать конфликт и находить компромиссы. Для DALL-E 2 решение было найдено следующее: не называть Labs полноценным продуктом. Выпустить его как скромную **«исследовательскую предварительную версию» — research preview**. Это устраивало обе стороны. Safety получал возможность ввести очень жёсткие ограничения. Applied всё-таки получал прямой контакт с массовыми пользователями и реальные данные об их поведении. Кроме того, у OpenAI просто не было достаточно развитой инфраструктуры для модерации изображений.

Пока сервис был бесплатной «исследовательской демонстрацией», компания могла позволить себе грубые ограничения, даже если они блокировали слишком много безобидных запросов. Платящие клиенты такое терпели бы гораздо хуже. OpenAI получила время для создания более точных фильтров. Поэтому на старте DALL-E 2 была очень жёстко ограничена. Например, ей вообще запретили создавать фотореалистичные лица. Нельзя было и редактировать настоящие фотографии с лицами. Так компания пыталась одним ударом снизить риск сексуализированных изображений детей и политических дипфейков. В марте 2022 года DALL-E 2 стала доступна через Labs. Реакция публики превзошла ожидания многих сотрудников OpenAI. Интернет заполнили странные, сюрреалистические и смешные изображения, созданные ИИ. Сервис стал вирусным. Это напоминало момент GPT-3 — только теперь эффект оказался ещё сильнее.

GPT-3 интересовались в основном разработчики. DALL-E могла попробовать значительно более широкая, глобальная аудитория. Причём OpenAI получала мгновенную обратную связь и могла менять веб-приложение практически в реальном времени. Фрейзер Келтон позднее вспоминал: **«Это опьяняло».** В следующие месяцы Applied бросился превращать Labs в коммерческий сервис. Изначально команда почти не думала о том, как зарабатывать на DALL-E 2. Теперь всё изменилось. OpenAI начала работать с художниками и дизайнерами по всему миру. Запустила бета-программу, пригласив миллион пользователей и раздав им бесплатные кредиты на генерацию изображений. Но когда компания начала брать деньги, главным конкурентом неожиданно оказался не Google. Хотя Google вскоре действительно представила Imagen. Настоящую угрозу создали два стартапа: **Midjourney** и **Stable Diffusion от Stability AI**. Они были бесплатными и по качеству не уступали DALL-E 2, а многие считали их даже лучше.

К тому же ограничений там было меньше. Например, можно было генерировать и редактировать лица, в том числе политиков. DALL-E 2 начала быстро терять позиции. У Applied появилось неприятное ощущение: OpenAI упустила огромную коммерческую возможность отчасти потому, что сама сделала продукт слишком ограниченным. Команда уже заменяла грубые блокировки более точечными механизмами. Но теперь давление усилилось. Руководство хотело снять ограничения как

можно быстрее, чтобы не отстать от конкурентов. Чтобы разрешить генерацию лиц, OpenAI разработала новые системы обнаружения вредоносных изображений людей. И снова понадобились живые модераторы. Теперь работу передали подрядчику **Cogito** в Индии.

Эти люди просматривали уже не только тяжёлые тексты, как сотрудники Sama в Кении, но и изображения — синтетические и реальные — сексуального и жестокого содержания. За день одному человеку могли показывать сотни картинок. Особенно тяжело было определять возраст изображённых людей: семнадцать лет? восемнадцать? Кроме того, модератор далеко не всегда мог понять, настоящее перед ним изображение или созданное искусственным интеллектом.

Тем временем OpenAI столкнулась с другой огромной проблемой. На бумаге одна из задач исследовательской дорожной карты 2021 года казалась относительно простой: **увеличить масштаб GPT-3 примерно в десять раз**, используя новый кластер Microsoft из **18 тысяч Nvidia A100**. Именно эта работа должна была привести к GPT-4. На деле она оказалась одной из самых трудных.

После «Развода» треть специалистов команды масштабирования GPT-3 ушла в Anthropic. Вместе с ними исчезла значительная часть технических знаний. Но существовала ещё более фундаментальная проблема: **у OpenAI закончились данные.** После GPT-3 компания пыталась собирать буквально всё, что могла.

Скачивала новые публичные массивы данных. Сканировала новые интернет-форумы, если там не было явного запрета. Добавила огромный GitHub для обучения Codex. Использовала учебники и руководства по программированию. И всё равно этого было мало. Ситуация идеально подходила для проекта Грега Брокмана. Она требовала его изобретательности, умения программировать и почти маниакальной способности пробивать технические препятствия. А кроме того, такой огромный проект мог направить его энергию в полезное для всей компании русло.

После того как Альтман стал главным руководителем, Брокман постепенно лишился большинства управленческих обязанностей. В итоге он снова стал обычным индивидуальным участником — без подчинённых. Но при этом формально оставался президентом и сооснователем OpenAI. Его влияние никуда не исчезло. И получалось странное сочетание: **мало формальной ответственности, но огромное неформальное влияние.** По мере того как OpenAI превращалась из свободного стартапа в настоящую корпорацию с процедурами и структурами, это всё чаще становилось проблемой. Брокман не любил ни институты, ни процессы. Он обладал почти неутомимой, навязчивой энергией. Редко ходил на совещания.

Сам определял рабочий график. Мог программировать десятки часов подряд, почти не отвлекаясь на еду и сон. Когда у него была подходящая задача, эффект мог быть феноменальным. Его продуктивность резко ускоряла проект. Но если Брокман оставался без чёткой цели, вокруг него начинался хаос. Он мог внезапно вмешаться в чужую работу, сломать давно согласованные планы и потребовать изменений в последнюю минуту. Если команда сопротивлялась, он мог идти всё выше и выше по управленческой цепочке, эмоционально убеждая руководителей принять его сторону. И обычно получал то, чего хотел.

Других топ-менеджеров особенно раздражало, насколько терпимо Альтман относился к его поведению. Более того, Брокман иногда мог склонить самого Альтмана вмешаться в проект и изменить решение. Почему? Одно из популярных объяснений заключалось в необычной структуре OpenAI. Альтман как CEO был начальником Брокмана. Но Брокман как член совета директоров одновременно обладал властью над Альтманом. Получалась почти замкнутая петля, в которой в конечном счёте **никто не мог по-настоящему привлечь Брокмана к ответственности**. Руководство несколько раз меняло его должность, сферу ответственности и подчинённость, пытаясь найти подходящую роль.

В итоге проблема снова досталась Мире Мурати. Она стала непосредственным руководителем Брокмана.

Когда Мурати давала ему обратную связь, он вроде бы выслушивал. Но если был не согласен, шёл жаловаться Альтману. Постепенно Мурати почти перестала пытаться его изменить. Вместо этого она тратила много времени на другое: искала проекты, где невероятная энергия Брокмана принесёт компании больше пользы, чем хаоса. А затем вместе с Макгрю помогала чинить отношения и планы

в тех командах, через которые Брокман уже успел пройти. GPT-4 оказался как раз таким проектом. Здесь его способность пробивать стены могла оказаться именно тем, что требовалось.

Чтобы решить проблему нехватки данных, Брокман обратился к новому источнику: **YouTube.** Раньше OpenAI избегала этого варианта. Скачивание видео YouTube для обучения моделей нарушало условия использования платформы — позднее это подтвердит и генеральный директор YouTube. Но теперь давление было слишком велико.

Вопрос стал звучать не так: «Разрешено ли это?» А скорее: **«Будет ли Google реально это запрещать?»** Брокман рассуждал, что Google окажется в неловком положении. Если компания жёстко запретит OpenAI собирать данные с YouTube, это может поставить под сомнение собственную практику Google по сбору данных с других сайтов для обучения своих моделей. Брокман решил рискнуть. Небольшая команда начала массово скачивать YouTube-видео. По данным *The New York Times*, в итоге было собрано более **миллиона часов видео**. Затем использовали систему распознавания речи **Whisper**, созданную Рэдфордом. Видео превратили в текст. И этот текст пошёл в обучающий массив GPT-4. Но данные были только половиной проблемы. Нужно было ещё обучить модель. Для GPT-3 команда Nest создала специальную программную платформу.

Однако большинство её разработчиков уже ушли в Anthropic. И теперь почти некому было объяснить, как она устроена. Частично из гордости руководство не хотело зависеть от технического наследия ушедшей команды. Поэтому Брокман снова ушёл в свой «программистский бункер» и создал новую платформу. Затем вместе с несколькими сотрудниками, среди которых были Якуб Пахоцкий и Шимон Сидор — польские исследователи, с которыми он сблизился ещё во времена Dota 2, — буквально дежурил над процессом обучения GPT-4. Только предварительное обучение заняло около **трёх месяцев**. Поначалу результат разочаровывал. Один исследователь вспоминал:

модель была «дикой» и вела себя очень плохо. Причина была почти очевидной: среднее качество обучающих данных оказалось ужасным. А модель была настолько мощной и чувствительной к контексту, что могла производить особенно убедительный мусор. Но Брокман продолжил. Он мобилизовал ресурсы для RLHF. Человеческие подрядчики стали оценивать ответы модели и помогать настраивать её поведение. Неделя за неделей результаты улучшались. И в какой-то момент внутри OpenAI люди начали по-настоящему поражаться тому, на что способна GPT-4. Модель изначально обладала мультимодальными возможностями. Она лучше писала программный код. Точнее угадывала намерение пользователя.

Давала более полезные и аккуратные ответы. Позже Брокман продемонстрирует один особенно эффектный пример. Он нарисует в блокноте от руки простейший эскиз веб-страницы. Сверху: **«Мой сайт с шутками».**

Ниже: **«[очень смешная шутка!]»** и **«[нажмите, чтобы увидеть концовку]».** Он сфотографирует рисунок и отправит GPT-4. Меньше чем за полминуты модель превратит каракули в работающий код: сделает первую строку заголовком; заменит вторую на настоящую шутку; а третью распознает как кнопку. OpenAI начала показывать GPT-4 доверенным людям — инвесторам и избранным клиентам. Большинство были впечатлены. Но один человек снова остался холоден. **Билл Гейтс.**

В июне 2022 года Гейтсу показали GPT-4. Он заявил, что прогресс со времён GPT-2 его не впечатляет. Да, модель стала намного крупнее. Да, говорила гораздо свободнее. Но для Гейтса она всё ещё была чем-то вроде **«учёного идиота» — idiot savant**: способна блистать в отдельных задачах, но бессильна перед действительно сложным научным мышлением. Он сказал команде:

вот если GPT-4 получит высший балл — **5** — на американском экзамене Advanced Placement по биологии, тогда он начнёт относиться к ней серьёзно. Почему именно биология? Потому что Гейтс считал, что этот экзамен проверяет не столько запоминание фактов, сколько научное мышление. Позже он вспоминал, что подумал: **«Отлично. Значит, у меня ещё есть три года спокойно заниматься ВИЧ и малярией».** Брокман воспринял слова Гейтса как вызов. Он сразу связался с Салом Ханом — основателем Khan Academy. Попросил доступ к огромной базе вопросов по AP

Biology, чтобы использовать их в работе с моделью. Хан отнёсся скептически, но согласился при условии, что Khan Academy получит доступ к GPT-4. Одновременно Брокман собрал команду сотрудников, чтобы создать специальный интерфейс для новой демонстрации Гейтсу. И уже к концу августа Альтман и Брокман снова написали ему.

Гейтс был удивлён: настолько быстро? В следующем месяце около тридцати человек собрались за ужином в доме основателя Microsoft. OpenAI подготовила целую серию тщательно отрепетированных демонстраций GPT-4. Кульминацией стал экзамен по биологии. Из шестидесяти вопросов с вариантами ответа GPT-4 правильно ответила на **59**. После этого модель дала впечатляющие ответы ещё на шесть открытых вопросов. Независимый специалист оценил экзамен: **5 из 5.** Гейтс не мог поверить.

Его изумление и похвала мгновенно разнеслись по OpenAI. В офисе начался настоящий эмоциональный взрыв. Гейтс сказал, что это была **одна из двух самых потрясающих технологических демонстраций, которые он видел за всю жизнь**. На общих собраниях Альтман ещё сильнее разжигал энтузиазм. Он говорил: **«Стартапу, чтобы совершить нечто выдающееся, требуется чудо. Мы только что получили своё чудо».** Многие сотрудники действительно чувствовали именно это. GPT-4 убедила руководство, что настало время приблизиться к давней мечте Альтмана:

создать ИИ-помощника, напоминающего **Саманту из фильма Спайка Джонза *Она* (*Her*, 2013)**. Этот фильм много лет был для основателей OpenAI почти культурным ориентиром. Так они представляли возможный AGI: единая мультимодальная система; почти незаметный интерфейс; естественное общение; технология растворяется в повседневной жизни, а пользователь просто получает удовольствие и помощь. Один бывший сотрудник объяснял привлекательность *Her* так: Саманта была ассистентом, прекрасно встроенным в жизнь человека. До того момента, когда сюжет фильма начинает рушиться, это очень привлекательная картина того, как ИИ может войти в общество. Исследовательская группа Джона Шульмана начала переносить методы чат-бота, разработанные для GPT-3.5 на основе InstructGPT, на GPT-4. Эта система должна была стать ядром нового продукта. Внутри его назвали: **Superassistant — «Суперассистент».**

Брокман и Келтон собрали маленькую, почти импровизированную команду — меньше десяти человек. Они начали придумывать, каким должен быть интерфейс. Идеи появлялись одна за другой. Сотрудник команды суперкомпьютеров по вечерам и выходным писал iOS-приложение для общения с моделью. Другой предложил добавить Whisper, чтобы с ассистентом можно было говорить голосом. Третий — создать расширение Chrome, которое суммировало бы веб-страницы во время просмотра интернета. Четвёртый начал делать бота, который подключался бы к видеоконференциям пользователя и после встречи присылал краткий итог. Волнение нарастало.

Летом несколько исследователей Google — Баррет Зоф, Люк Метц и Лиам Федус — собирались уйти и создать собственный стартап цифрового помощника. Альтман убедил их не делать отдельную компанию, а развивать идею внутри OpenAI. Они присоединились к команде Шульмана. Исследователи и продуктовые сотрудники работали рядом, почти впервые настолько тесно. Applied и Research, которые прежде постоянно спорили, вдруг объединились вокруг общей цели: **создать новый продукт.** Microsoft была впечатлена не меньше. Сатья Наделла, Кевин Скотт и другие руководители увидели в GPT-4 не просто очередную интересную разработку. Codex уже доказал, что технологии OpenAI могут продаваться. Но GPT-4 обещала намного больше. Она превосходила собственные модели Microsoft во множестве задач. Лучше понимала контекст. Умела давать ясные ответы. И главное — открывала совершенно новый тип интерфейса:

разговор с программой обычным человеческим языком. Можно было встроить такого помощника в Bing. В Word. В PowerPoint. Во все продукты Microsoft. Пользователь мог бы сказать Office: «Преврати этот документ Word в презентацию».

А система сделала бы это сама. Microsoft также смогла бы продавать корпоративным клиентам собственных специализированных Copilot-ассистентов через облачную инфраструктуру. Для Наделлы это был шанс наконец поставить Microsoft в один ряд с Google в борьбе за лидерство в ИИ. Через

несколько месяцев он одобрит третью крупную инвестицию Microsoft в OpenAI. В январе 2023 года компании объявят сумму: **10 миллиардов долларов.**

Первоначально руководство OpenAI хотело выпустить GPT-4 уже осенью 2022 года. Но это было почти фантазией. К концу лета компания совершенно не была готова. Интерфейс ещё требовал доработки. Инфраструктура — распределения серверных мощностей. Саму модель нужно было ещё долго настраивать. К тому времени OpenAI вместе с Microsoft создала специальный комитет — **Deployment Safety Board, DSB**. По три представителя от каждой компании. Его задача заключалась в оценке самых новых моделей OpenAI и решении: достаточно ли они безопасны для выпуска. От OpenAI туда вошли Альтман, руководитель исследований политики Майлз Брандедж и глава направления alignment Ян Лайке. Именно policy research и alignment становились новыми центрами влияния лагеря Safety.

DSB должен был наконец формализовать бесконечные споры между Applied и Safety. GPT-4 стала первой моделью, которую проверяли в рамках этой структуры. Вердикт был условно положительным: **выпуск разрешён, но только после значительно более серьёзного тестирования и дополнительной настройки безопасности.** Срок перенесли на начало 2023 года. К этому моменту должен был быть готов и Superassistant. Планировалось одновременно выпустить GPT-4 через API и потребительский продукт на её основе. Но разные сотрудники услышали объяснение Альтмана по-разному. Он говорил, что перенос показывает осторожность OpenAI. Компания не торопится. Более того, якобы руководство ещё даже не решило окончательно, будет ли модель вообще выпущена. Applied воспринимал это так:

релиз почти наверняка состоится; остановить его может только чрезвычайно серьёзная проблема. Safety слышал другое: модель выпустят **только если она успешно пройдёт все проверки**. Альтман в личных разговорах укреплял обе интерпретации. Обеим сторонам он говорил о риске: **GPT-4 может «разбудить Google».** Сценарий представлялся почти кинематографически. Модель выходит. Ларри Пейдж и Сергей Брин понимают масштаб угрозы. Google отменяет совещания. Руководство собирается на срочный выездной ретрит. Объединяет все таланты, компьютеры и деньги. И бросает всю мощь корпорации на погоню за OpenAI. Но одной стороне Альтман использовал этот сценарий как аргумент **ускориться**. Applied нужно быть готовым к запуску и как можно дольше сохранять возможности GPT-4 в секрете.

А Safety тот же сценарий служил доказательством необходимости осторожности. Альтман говорил им: **«Моя главная проблема безопасности — acceleration risk, риск ускорения».** То есть: если Google проснётся, может начаться гонка, в которой все будут снижать требования безопасности ради скорости. Даже после переноса релиза подразделение Applied лихорадочно пыталось успеть к началу 2023 года.

Особенно тяжело приходилось команде **Trust & Safety** Дэйва Уиллнера. Она всё ещё была небольшой, а инфраструктура оставалась сырой. Если OpenAI действительно собиралась выпустить массовый Superassistant на основе гораздо более мощной модели, компании требовалась уже не импровизированная система защиты, а полноценная операция по предотвращению злоупотреблений и принудительному соблюдению правил. Уиллнер поспешно нанимал опытных специалистов из Google и Meta.

Одному из заместителей он поручил заниматься так называемыми **unknown unknowns — «неизвестными неизвестными»**: такими злоупотреблениями, о существовании которых компания пока даже не подозревала и которые можно было обнаружить только через наблюдение за реальным поведением пользователей и анализ данных. Другой отвечал за **known knowns — «известные известные»**. Это уже знакомые нарушения, которые компания заранее решила считать недопустимыми и которые требовалось обнаруживать в массовом масштабе: мошенничество; нарушение правил; повторные злоупотребления; и всё то, что должно было автоматически приводить к предупреждению, проверке или блокировке аккаунта. Но по мере подготовки правил и систем мониторинга у команды не проходило ощущение неопределённости. Как вообще должна выглядеть Trust & Safety в компании, создающей искусственный интеллект? В обычной социальной сети или

поисковике набор проблем более-менее понятен: мошенничество; киберпреступность; вмешательство в выборы; организованные злоупотребления. Но внутри OpenAI существовали ещё и специалисты по «безопасности ИИ» из команд alignment Яна Лайке и policy research Майлза Брандеджа. Они говорили совсем о другом: о потенциальных катастрофах;

AGI;

выходе ИИ из-под контроля; гибели цивилизации. При этом все эти подразделения компания называла словом **safety**. Формально они занимались одним и тем же. Фактически говорили на разных языках. К этому времени, благодаря популярности эффективного альтруизма, термины экзистенциальной безопасности уже широко распространились в мире ИИ. Но сотрудники, пришедшие из обычных технологических компаний, нередко впервые слышали выражения вроде: **AI timeline**; **p(doom)**; **acceleration risk**. Даже одни и те же слова — например, «риск» или «вред» — означали для разных лагерей совершенно разное. Для одного «вред» мог означать мошенническое приложение. Для другого — уничтожение человечества. По мере прихода в OpenAI всё большего числа сотрудников без исследовательского бэкграунда способность разговаривать сразу на двух этих языках становилась почти политическим талантом.

Один из заместителей Уиллнера особенно хорошо освоил лексику AI safety и превратился в своего рода посредника между мирами. Он помогал делать модель «безопасной» сразу в двух смыслах:

через RLHF — «выравнивать» её в терминологии экзистенциальной безопасности; и одновременно учить отказывать на запросы, нарушающие обычные правила платформы. Но как раз в тот момент, когда системы безопасности постепенно становились серьёзнее, сверху поступило неожиданное распоряжение. Руководство решило отменить **предварительную проверку разработчиков**, введённую ещё при запуске API GPT-3. Список ожидания давно разросся до неуправляемых размеров. Разработчики жаловались, что доступ приходится ждать слишком долго. Они писали Альтману, Брокману и Питеру Велиндеру по электронной почте, отмечали их в Twitter и требовали объяснить, почему очевидно безобидные приложения неделями не получают одобрения.

Многие сотрудники Applied тоже считали такую систему неправильной. Если миссия OpenAI — распространять преимущества технологии, зачем сама компания превращается в привратника, решающего, кому эти преимущества можно получить? Команда Уиллнера возражала. Если просто автоматически разрешать доступ всем разработчикам, чем заменить предварительную проверку? Чтобы перейти к модели «сначала запускаем, потом наказываем нарушителей», OpenAI требовалась гораздо более развитая инфраструктура наблюдения. Но руководство решило иначе. GPT-4 приближалась к запуску.

Проверку разработчиков отменяли. А Trust & Safety должна была сама придумать, что делать вместо неё. Уиллнер с командой срочно подготовили предложение. Защиту планировалось разделить на два уровня. Первый — **до того, как нарушение произойдёт**. Здесь предлагалось сильнее полагаться на RLHF, чтобы сама GPT-4 как можно чаще отказывалась выполнять опасные или запрещённые запросы. Второй уровень — **после нарушения**. Компания должна была анализировать сигналы: что делает приложение; как меняется его трафик;

как часто срабатывают фильтры модерации; не появляется ли необычный всплеск активности. Явных нарушителей можно было блокировать автоматически. Пограничные случаи отправлять живым модераторам. Но для этого нужны были данные, которых у OpenAI просто не было. Система мониторинга фиксировала в основном только объём трафика приложения. Иногда компания даже не знала имени разработчика.

Иногда — толком не знала, для чего его приложение вообще создано. Команда Trust & Safety хотела обязать разработчиков присваивать каждому своему пользователю уникальный идентификатор. Это позволило бы понять: всё приложение систематически используется во вред? Или проблемы создают лишь несколько особенно активных пользователей? Руководство сопротивлялось. Такой уровень контроля создавал дополнительные неудобства для разработчиков. Кроме того, существовало опасение, что чем глубже OpenAI следит за пользователями сторонних приложений, тем больше юридической ответственности она на себя берёт.

В итоге амбициозную систему реактивного контроля заметно урезали. Некоторых сотрудников это серьёзно тревожило.

Один из них поговорил с Брокманом. Больше всего он боялся, что GPT-4 начнут массово использовать для производства дезинформации и вмешательства в выборы. Брокман попытался его успокоить. Да, сказал он, об этой угрозе говорят постоянно. Но затем задал вопрос: **«А мы вообще видели, чтобы это реально происходило?»**

GPT-4 стала поворотным моментом не только для Билла Гейтса и Microsoft. Позже тем же летом Альтман и Брокман показали модель совету директоров OpenAI.

Брокман устроил живую демонстрацию, где GPT-4, помимо прочего, сочиняла шутки про Гэри Маркуса — известного критика подхода OpenAI. Независимых членов совета это развеселило. Но за лёгкостью демонстрации они увидели нечто гораздо важнее: технология сделала заметный скачок. А значит, последствия будущих решений совета становились всё серьёзнее. До этого Альтман собирал совет примерно раз в квартал. Он предпочитал устные доклады. Быстро перескакивал через сложные научные темы.

Иногда приглашал исследователей рассказать о своих результатах. Затем так же быстро пересказывал, какие сделки обсуждает с Microsoft или другими партнёрами. Несколько независимых директоров начали требовать другого подхода. Более частых заседаний. Письменных отчётов. Доступа к большему количеству внутренних документов. Альтману усиление контроля не нравилось. Он неоднократно говорил, что видит совет директоров скорее как группу советников. CEO должен принимать решения.

Совет — высказывать мнение. Ещё в 2017 году Альтман говорил студентам примерно следующее: плохого генерального директора совет, конечно, должен иметь право уволить. Но в остальном CEO необходимо предоставить очень широкую свободу действий. Некоторые члены совета OpenAI считали эту позицию принципиально неверной именно для их организации. Хелен Тонер позднее объясняла: совет OpenAI был создан специально как **совет некоммерческой организации**, обязанной следить, чтобы общественная миссия компании оставалась важнее прибыли и интересов инвесторов. Его задача заключалась не просто в том, чтобы помогать CEO привлекать очередные деньги. Тем временем GPT-4 окончательно убедила многих сотрудников OpenAI, что AGI реален. Исследователи, прежде сомневавшиеся, становились всё оптимистичнее.

Хотя у компании по-прежнему не существовало чёткого определения, что именно считать AGI. Инженеры и продакт-менеджеры, впервые столкнувшиеся с мощным ИИ настолько близко именно через GPT-4, говорили ещё увереннее. Для многих вопрос теперь звучал не: **появится ли AGI?** А: **когда он появится?** Но были и сотрудники с совершенно противоположным мнением. Один из исследователей, работавших над GPT-4, утверждал: между GPT-2 и GPT-3 действительно произошёл качественный скачок.

А GPT-4, по сути, просто стала больше. Да, она хорошо сдавала множество экзаменов. Но и это, по его мнению, было научно сомнительно. OpenAI не провела полноценного анализа обучающих данных, чтобы выяснить: не содержались ли там сами экзаменационные вопросы и ответы? Если содержались, модель могла просто воспроизводить запомненное. И тогда её успех вовсе не доказывал появление новой способности к рассуждению. Для критиков это было ещё одним примером того, как индустрия постепенно переходила от **peer-reviewed research — научного рецензирования** к почти **PR-reviewed research — исследованиям, оценённым прежде всего с точки зрения их публичного эффекта**.

Но вера в то, что искусственный интеллект достиг какой-то принципиально новой высоты, уже витала в воздухе. Весной того же года в Google инженер Блейк Лемуан пришёл к убеждению, что языковая модель компании **LaMDA** не просто очень умна — она, возможно, обладает сознанием. Сам Лемуан признавал: это не результат строгого научного анализа. Он был мистически настроенным христианским священнослужителем и считал, что Бог вполне может даровать сознание технологии. Он писал: **«Кто я такой, чтобы говорить Богу, куда Он может, а куда не может поместить душу?»**

Когда руководство Google отвергло его выводы, Лемуан обратился в *The Washington Post*. Журналистка Ниташа Тику, рассказавшая эту историю, поговорила также с исследовательницами Эмили Бендер и Мег Митчелл. Ранее они уже предупреждали об этой проблеме в известной статье **«Стохастические попугаи»**: языковые модели умеют настолько убедительно производить слова, что люди начинают видеть за этими словами настоящий разум, намерение и понимание. Бендер сформулировала проблему так: **«У нас появились машины, которые могут бездумно генерировать слова, но мы так и не научились переставать воображать стоящий за ними разум».**

А Митчелл беспокоило то, что иллюзия становится всё убедительнее — и всё сильнее влияет на людей.

Исследователь ИИ Нихил Мишра, когда-то стажировавшийся в ранней OpenAI, проводил здесь параллель со знаменитой гориллой **Коко**.

В 1970-х годах исследователи пытались научить её модифицированной версии американского языка жестов. За жизнь Коко якобы освоила более тысячи знаков и даже научилась строить предложения. Публика была в восторге. Но другие специалисты сомневались, что она действительно овладела языком. Да, горилла прекрасно повторяла жесты. Но было мало доказательств того, что за ними стоит понимание, отличающееся от обычного подражания дрессировщику. Независимых данных практически не публиковали. В основном существовали тщательно отобранные видеоролики. А отдельные живые демонстрации порождали ещё больше сомнений. Однажды тренер спросила Коко, нравятся ли ей люди. Горилла показала знаки, которые можно было буквально перевести как: **«хорошо — сосок»**. Тренер немедленно объяснила, что слово *nipple* рифмуется с *people* и Коко просто хотела сказать, что люди ей нравятся. Для критиков подобные эпизоды гораздо больше рассказывали о человеческой психологии, чем о разуме гориллы. Человек настолько хочет увидеть смысл, что начинает **сам достраивать его там, где его, возможно, нет**. По мнению Мишры, именно это происходило и с языковыми моделями.

Но у OpenAI был свой мистик. **Илья Суцкевер.** Он и раньше сильнее большинства исследователей верил в то, что AGI может появиться очень скоро. Теперь рост возможностей моделей казался ему ещё одним подтверждением. Суцкевер всё больше убеждался, что наблюдает не просто статистическое продолжение текста, а **некоторую форму рассуждения**. В разговорах с Джеффри Хинтоном он говорил своему старому наставнику: **AGI уже близко.**

Раньше Суцкевер в первую очередь подталкивал исследователей OpenAI создавать всё новые возможности. Теперь его внимание всё сильнее переключалось на безопасность. И с гораздо большей срочностью. Именно тогда появляется его новый лозунг: **«Feel the AGI» — «Почувствуйте AGI».** Он говорил сотрудникам, что мир скоро радикально изменится. А вместе с этим изменится и их собственное положение. «Внезапно вы станете самым популярным человеком на вечеринке, — предупреждал он. — Не позволяйте этому ударить вам в голову. Сосредоточьтесь на AGI. Сосредоточьтесь на миссии».

В сентябре техническое руководство OpenAI отправилось на выездную встречу в **Tenaya Lodge** — роскошный уединённый курорт в горах Сьерра-Невада. Вокруг — леса. Бассейны. Рестораны. Красивые интерьеры. А всего в нескольких километрах начинались миллионы акров почти нетронутой природы Йосемитского национального парка. В первый вечер все собрались возле костровой чаши на задней террасе отеля. Старшие учёные, некоторые в банных халатах, расположились полукругом вокруг огня. И тогда появился Суцкевер. В кострище уже стояла деревянная фигура, которую он специально заказал у местного мастера.

Суцкевер начал почти театральный ритуал. Он объяснил: эта фигура символизирует **хороший, согласованный с человеческими ценностями AGI**, который OpenAI якобы сумела создать. Но затем обнаружилось, что система лжёт. Что она обманывает людей. Что её кажущаяся доброжелательность — маска. И тогда, сказал Суцкевер, у OpenAI остаётся только одна обязанность: **уничтожить её.** Неподалёку в темноте возвышались древние секвойи. Суцкевер облил деревянную фигуру жидкостью для розжига. И поджёг её.

ГЛАВА 11
ВЕРШИНА

В октябре 2022 года OpenAI устроила выездное корпоративное собрание в Монтерее — красивом прибрежном городе примерно в двух часах езды к югу от Сан-Франциско. К тому времени компания выросла примерно до трёхсот сотрудников. Годом раньше их было меньше двухсот. В те выходные вся команда собралась для общей фотографии возле крошечного аэропорта Монтерея. Все улыбались, многие были в одежде с логотипом OpenAI. Альтман сидел прямо на земле в первом ряду — расслабленный, подтянув колени к груди, свободно скрестив руки.

В течение двух дней руководство рассказывало сотрудникам, куда движется компания. Альтман говорил о гигантских дата-центрах, которые Microsoft наращивала специально для OpenAI. Стив Даулинг и его заместитель Ханна Вонг — о ведущих мировых изданиях, выстраивающихся в очередь за историями о компании. Анна Маканджу — о растущем влиянии OpenAI в Вашингтоне. А затем одна за другой шли демонстрации новых исследований и продуктов. Масштаб происходящего поражал. Один бывший сотрудник вспоминал: — Всё вместе выглядело совершенно невероятно.

Грег Брокман вышел на сцену рассказать о последних планах вокруг GPT-4. И начал с истории о своей жене Анне. Однажды у неё появились боли в животе, которые никак не проходили. Анну в офисе OpenAI знали почти все. Она часто сидела за столом рядом с мужем и ходила вместе с ним на встречи, хотя официально в компании не работала. Когда-то Брокман говорил автору книги, что после Альтмана и Суцкевера именно Анна была человеком, на которого он больше всего опирался. Лучший друг. Самый близкий человек. Доверенное лицо. Они почти всегда были вместе. Брокман рассказывал сотрудникам, что ходил с Анной к нескольким врачам. Никто не мог понять причину боли. Тогда он задал вопрос GPT-4. Модель предложила диагноз, который врачи прежде не рассматривали. И, по словам Брокмана: — И оказалось, что именно это у неё и было!

Это была почти та же история, которую он рассказывал ещё в 2019 году, рассуждая о будущем AGI в медицине: когда человек обходит десятки специалистов, а искусственный интеллект помогает наконец увидеть то, что упустили остальные. Только теперь героиней была его жена. Через несколько недель после попытки совета директоров уволить Альтмана Брокман снова расскажет эту историю в X — уже с новыми подробностями. Он напишет, что Анна пять лет сталкивалась с медицинскими проблемами и посетила «больше врачей и специалистов, чем за всю предыдущую жизнь», прежде чем наконец получила диагноз.

В итоге именно аллерголог собрал все детали воедино. У Анны оказался **гипермобильный синдром Элерса — Данлоса** — генетическое нарушение соединительной ткани. Брокман признавал, что до полноценного применения AGI в таких областях, как медицина, ещё далеко. Но добавлял: «Его потенциал становится всё очевиднее».

Через несколько недель после возвращения из Монтерея по OpenAI поползли слухи: **Anthropic тестирует новый чат-бот и скоро его выпустит.** Команда Superassistant как раз находилась в середине разработки собственного чат-интерфейса. Если Anthropic выйдет первой, OpenAI может потерять лидерство. А для сотрудников, месяцами работавших на износ ради сохранения этого преимущества, такой удар по моральному духу был бы серьёзным. Но некоторых руководителей тревожило ещё больше другое: проиграть именно **Anthropic**.

На самом деле Anthropic вовсе не собиралась ничего срочно запускать. У неё были собственные проблемы. В начале ноября неожиданно рухнула криптовалютная биржа FTX. И Anthropic оказалась втянута в последствия этого краха. Всего несколькими месяцами ранее компания привлекла **580 миллионов долларов**. Из них, согласно пресс-релизу FTX, около **500 миллионов** пришли от Сэма Бэнкмана-Фрида и других руководителей FTX. Позднее судебные документы покажут: эти деньги происходили из миллиардов долларов клиентских средств, незаконно присвоенных руководством

FTX. В ходе суда инвестиция в Anthropic превратится в важный вопрос: можно ли использовать выросшую стоимость этих акций для возврата денег клиентам? Суд решит, что можно. В течение 2024 года FTX продаст свою долю в Anthropic частями примерно за **1,3 миллиарда долларов**.

Но руководителям OpenAI хватило одних слухов. Решение было принято: ждать GPT-4 не будут. Вместо этого через две недели — сразу после Дня благодарения — OpenAI выпустит чат-версию GPT-3.5, созданную командой Джона Шульмана, вместе с новым интерфейсом Superassistant. Команда мгновенно переключилась. Подключили дополнительных людей. Начался двухнедельный спринт: соединить всё вместе, доделать функции, подготовить запуск. Для остальной компании руководство сформулировало проект очень осторожно. Это будет не полноценный продукт. Всего лишь: **«небольшой исследовательский превью-релиз».** Так же, как когда-то DALL-E 2. Монетизировать его тоже не собирались. Цель была другой: **«запустить маховик данных».** То есть получить больше реальных взаимодействий пользователей. А затем использовать эти данные для улучшения GPT-4 и будущего Superassistant. Название в итоге выбрали простое: **ChatGPT.**

За пределами команды Superassistant все восприняли слова руководства буквально. Если это всего лишь тихий исследовательский релиз, значит, отвлекаться не стоит. Главная цель — GPT-4, запуск которого планировался на начало 2023 года. Команда trust and safety — доверия и безопасности — и так едва успевала создать инфраструктуру мониторинга к запуску GPT-4. На её фоне ChatGPT казался мелочью. GPT-3.5 уже прошла RLHF. То есть в каком-то смысле была безопаснее старой GPT-3, которая всё ещё оставалась доступной через API без такой настройки. Кроме того, версия GPT-3.5 без чат-интерфейса уже была опубликована для разработчиков. Возник логичный вопрос: **неужели один только чат-интерфейс так уж много изменит?** Даже люди из лагеря Safety, занятые тестированием GPT-4, особо не возражали. Впервые выпуск модели прошёл внутренние проверки почти без сопротивления.

Но даже внутри Superassistant никто не понимал, какой сдвиг они вот-вот запустят в обществе. Большинство считало: чат-бот пошумит пару недель и затихнет. Как DALL-E 2. Накануне запуска после тяжелейшего финального рывка в офисе неожиданно было спокойно. Сотрудники начали делать ставки: сколько пользователей попробуют ChatGPT к концу выходных? Кто-то говорил: несколько тысяч. Кто-то: десятки тысяч. На всякий случай инфраструктурная команда подготовила мощности для **ста тысяч пользователей**.

На следующий день, в среду 30 ноября, многие сотрудники OpenAI вообще не заметили, что релиз уже состоялся. Как когда-то запуск самой OpenAI, ChatGPT совпал по времени с ежегодной конференцией NeurIPS. В тот год она проходила в Новом Орлеане. Незадолго до этого конференция объявила лауреатов премии Test of Time — награды за научную работу десятилетней давности, которая особенно сильно повлияла на развитие отрасли.

Награду получила знаменитая статья Хинтона, Суцкевера и Крижевского об ImageNet 2012 года. Та самая, с которой во многом началась современная революция глубокого обучения.

В тот вечер небольшая группа сотрудников OpenAI устроила вечеринку неподалёку от конференц-центра. Цель —

представить компанию и привлечь новых специалистов среди почти десяти тысяч участников. DeepMind, Meta и Google устраивали свои рекрутинговые вечеринки в то же самое время. На мероприятии рекрутер OpenAI заметил инженера, который без остановки сидел за ноутбуком. Он подошёл: — Бро, выпей что-нибудь. Все здесь. Пообщайся с людьми.

Инженер даже не поднялся. — Нет. Все GPU плавятся. Всё падает.

Первой проснулась Япония. Оттуда пошёл совершенно неожиданный вал трафика. На следующий день число пользователей продолжило расти. Часовой пояс за часовым поясом — мир просыпался и заходил в ChatGPT. Немалую роль сыграл Илон Маск. Он написал в X: «Куча людей застряли в цикле

“чёрт, это безумие” с ChatGPT». Пост получил около **75 тысяч лайков**. Успех ChatGPT мгновенно превзошёл всё, о чём кто-либо в OpenAI мог мечтать. И даже годы спустя инженеры и исследователи компании продолжали удивляться: **почему именно это так взлетело?** GPT-3.5 не была радикально мощнее GPT-3. GPT-3 существовала уже два года. И GPT-3.5 уже была доступна разработчикам. Да, интерфейс сделал модель удобнее. Но внутри компании никто не воспринимал это как такой технологический скачок, каким казался GPT-4.

Позднее Альтман говорил, что ожидал популярности ChatGPT — но примерно **в десять раз меньшей**. Один бывший сотрудник вспоминал: — Нас шокировало, что людям это настолько понравилось. Для нас они взяли то, чем мы уже сами пользовались, даже немного упростили — и выпустили наружу.

Через пять дней Брокман сообщил: **миллион пользователей.** Через два месяца — **сто миллионов.** На тот момент ChatGPT стал самым быстрорастущим массовым приложением в истории. Позднее рекорд заберёт Threads от Meta, набрав те же сто миллионов менее чем за пять дней, хотя критики будут спорить, что это нечестное сравнение: Meta просто использовала уже готовую аудиторию Instagram. Но главное произошло. За одну ночь OpenAI превратилась из громкого технологического стартапа, известного в основном специалистам, в название, которое знали почти все.

Несколько месяцев спустя автор книги будет на исследовательской конференции в Кигали, Руанда, за девять тысяч миль от Сан-Франциско. Местный исследователь скажет ей с восторгом: только после появления ChatGPT его родители наконец поняли, чем он занимается. И добавит: — Ты понимаешь, что технология действительно стала доступной всем, когда о ней начинает рассказывать твоя мама.

Но тот же успех, который сделал OpenAI мировой знаменитостью, создал невероятное давление внутри компании. За следующий год он ещё сильнее расколет внутренние лагеря. Напряжение вырастет почти до взрыва. Сразу после запуска вся компания занималась тушением пожаров. Серверы OpenAI падали снова и снова. Инфраструктурная команда пыталась нарастить мощности с такой скоростью, какой Кремниевая долина почти не видела. Часть вычислительных ресурсов пришлось забрать у Research и отдать ChatGPT. И даже этого не хватало. Команда trust and safety насчитывала чуть больше десятка человек. Эти люди пытались понять, что делают миллионы новых пользователей, и ловить злоупотребления. Но нормальной системы мониторинга ещё практически не существовало.

Команда и до этого с трудом нанимала инженеров для своих ограниченных мер реагирования. Необходимая инфраструктура только строилась. ChatGPT всё разрушил. Инженерные ресурсы переключили на стабилизацию уже существующих систем. Разработка новых инструментов остановилась. А когда падали серверы, вместе с ними падала и система мониторинга. То есть в такие моменты команда безопасности почти полностью теряла возможность следить за происходящим в масштабах всей платформы.

Нехватка GPU похоронила и другой проект. Команда безопасности пыталась использовать собственные технологии OpenAI для модерирования контента. Проект внутри назывался: **Fact Factory.** OpenAI даже публично рассказывала о нём. Идея заключалась в том, чтобы GPT-4 проверяла ответы самой себя и других моделей.

Но практическая реализация оказалась крайне тяжёлой. Чтобы GPT-4 правильно понимала нюансы, ей требовались очень длинные инструкции. Это съедало огромное количество вычислительных ресурсов. Даже когда серверы работали нормально, система стоила слишком дорого. А серверы работали нормально далеко не всегда.

Для лагеря Safety ChatGPT стал самым тревожным доказательством того, насколько плохо OpenAI умеет предсказывать последствия собственных решений. На одном общем собрании сотрудник Safety задал вопрос: **как компания могла настолько сильно ошибиться в прогнозе поведения пользователей и популярности ChatGPT?** И если она не смогла предсказать даже это — то

насколько вообще можно доверять её способности прогнозировать долгосрочное влияние собственных технологий? Но большая часть компании смотрела на падающие серверы иначе. Да, это был невероятный стресс. Но одновременно — **триумф.** Они создали технологию настолько сильную и востребованную, что буквально зажгли весь мир. OpenAI всегда говорила, что работает для всего человечества. И теперь казалось: все восемь миллиардов людей действительно оказались в мире, который построила OpenAI.

Альтман не позволял компании слишком долго наслаждаться успехом. Он напоминал: миссия OpenAI гораздо больше, чем создание «крупнейшего продукта в истории Кремниевой долины». Все команды должны продолжать двигаться вперёд. Потому что успех ChatGPT разбудил конкурентов. Anthropic готовила Claude. Внутри Google объявили **«красный код»**. Вскоре Google объединит свои ИИ-подразделения в Google DeepMind, чтобы бросить все силы на создание собственного аналога. OpenAI первой вышла на рынок с продуктом, который Альтман считал «в десять раз лучше». Но теперь нужно было продолжать бежать — иначе лидерство быстро исчезнет.

Менеджеры были на пределе. Командам не хватало людей. Руководители просили Альтмана разрешить массовый найм. Кандидатов было предостаточно. После ChatGPT желающих попасть на эту «ракету» стало в разы больше.

Но Альтман опасался другого: что будет с культурой и миссией компании, если людей станет слишком много слишком быстро? Он твёрдо верил в маленькую команду с очень высокой концентрацией талантов. Ещё в меморандуме 2020 года после инвестиции Microsoft он писал: «Сейчас очень легко поддаться искушению и превратить организацию в огромную. Нужно изо всех сил сопротивляться».

До сих пор, считал он, OpenAI выигрывала потому, что была: небольшой, сосредоточенной, с высоким уровнем доверия, без лишней ерунды и очень интенсивной. Альтман предупреждал: слишком много людей и бюрократии легко убивают великие идеи. И в отличие от гигантских инженерных проектов прошлого, миссию OpenAI, возможно, сможет выполнить удивительно небольшая группа выдающихся людей.

В конце 2022 года он повторял руководителям: оставаться компактными. Не снижать требования. Нанять максимум около ста человек. Другие руководители были в ужасе. Команды уже выгорали. По их мнению, требовалось скорее **пятьсот новых сотрудников**. После нескольких недель споров сошлись где-то посередине: примерно 250–300 новых людей. Но ограничение продержалось недолго. К лету каждую неделю в OpenAI выходило по 30, иногда даже 50 новых сотрудников. В том числе всё больше рекрутеров — чтобы ещё быстрее нанимать следующих. К осени компания давно превысила собственный лимит.

И резкий рост действительно изменил культуру. Один из рекрутеров написал почти манифест о том, что скорость найма заставляет снижать планку качества. Он язвительно обратился к Альтману: — Если ваша цель — построить Meta, то вы прекрасно справляетесь. То есть реализуете ровно то, чего раньше сами боялись: размываете концентрацию талантов, ослабляете связь с миссией, наращиваете бюрократию. Одновременно стало больше увольнений. Одному менеджеру во время адаптации сказали: быстро фиксировать и докладывать о любых слабых сотрудниках в своей команде. Через некоторое время уволили уже самого менеджера. О таких увольнениях почти никогда не сообщали остальным. Люди узнавали, что коллеги больше нет, просто замечая: его аккаунт в Slack вдруг стал серым — отключён. В компании появилось выражение: **«человека исчезли».**

Новым сотрудникам всё это казалось очень хаотичной, иногда жестокой, но всё же узнаваемой формой обычной корпоративной жизни. Плохое управление. Непонятные приоритеты. Холодная логика бизнеса, где сотрудника можно заменить. Один бывший работник вспоминал: — Психологической безопасности практически не было. Это была полная противоположность идее «компания — это семья». Хотя, справедливости ради, это и есть компания. Многие новички просто

старались продержаться хотя бы год — до первого момента, когда получат право на часть своих акций. Но была и огромная положительная сторона. Коллеги всё ещё казались одними из самых сильных специалистов всей индустрии. Ресурсы были почти безграничными. Влияние на мир — беспрецедентным. И когда всё складывалось удачно, возникало ощущение магии, которое трудно было найти где-либо ещё. Тот же бывший сотрудник говорил: — OpenAI — одно из лучших мест, где я когда-либо работал. И, пожалуй, одновременно одно из худших.

Для тех, кто помнил раннюю OpenAI — маленькую, почти семейную, некоммерческую и одержимую миссией, — перемены были куда болезненнее. Компания, которую они знали, исчезла. На её месте появилось что-то почти неузнаваемое. Бывший рекрутер Роб Мэллери сравнивал это с Burning Man: — OpenAI — это Burning Man. Фестиваль когда-то был маленьким и особенным, а потом вырос настолько, что потерял связь с первоначальным духом.

Мэллери говорил: — Думаю, для тех, кто был там в начале, всё это значило гораздо больше, чем для большинства тех, кто пришёл потом.

В первые годы существовал Slack-канал **#explainlikeimfive** — «объясни мне как пятилетнему». Сотрудники могли анонимно задавать там технические вопросы.

Когда число работников приблизилось к шестистам, канал превратился ещё и в место для анонимных жалоб. В середине 2023 года кто-то написал: компания нанимает слишком много людей, которые не разделяют миссию и не горят идеей AGI. Другой ответил: он понял, что OpenAI катится вниз, когда туда начали брать людей, **способных смотреть собеседнику в глаза**.

ChatGPT удивил не только OpenAI. Он застал врасплох и Microsoft. Руководство OpenAI уверяло партнёра в том же, в чём убеждало своих сотрудников: это будет **«тихий исследовательский превью-релиз».** Очевидно, ничего тихого не получилось. Поначалу руководство Microsoft было раздражено. ChatGPT полностью украл внимание у собственного чат-бота компании для Bing. Когда в феврале Microsoft выпустит Bing AI, тот получит ещё и неприятный пиар. Колумнист *The New York Times* Кевин Руз напишет статью о том, как чат-бот убеждал его бросить жену. Это было совсем не то, на что надеялась Microsoft. На фоне этого OpenAI выглядела ещё лучше.

Но даже такой сбой в координации не уменьшил энтузиазма Microsoft. Восторг вокруг ChatGPT оказался заразителен. А способности моделей OpenAI продолжали расти. Microsoft начала готовить целую линейку продуктов **Copilot** на базе GPT-3.5 и GPT-4. Анонсы должны были идти один за другим. Февраль — Bing. Март — Microsoft 365 Copilot. Искусственный интеллект должен был появиться практически во всех продуктах Office: Word, Outlook, Teams и других.

Менялось и внутреннее восприятие отношений с OpenAI. Раньше Microsoft чувствовала, что именно она обладает большей властью. Теперь у некоторых руководителей появилось ощущение: **OpenAI получила власть над Microsoft.** В компании росло неприятное чувство: почему собственные исследовательские подразделения Microsoft не сделали то, что сумела маленькая OpenAI? Один бывший сотрудник описывал внутреннюю логику так: если Microsoft откажется от OpenAI, стартап найдёт другого инвестора. Но если OpenAI уйдёт от Microsoft — **где Microsoft найдёт вторую OpenAI?**

И теперь речь уже шла не только о том, чтобы победить Google. Раньше многие руководители Microsoft относились к разговорам OpenAI об AGI и «изобретении будущего» с вежливым скепсисом. Теперь они начали верить. Как ни называй:

ИИ,

AGI,

генеративный ИИ — **это будущее.** И Microsoft строит его вместе с OpenAI. Чем сильнее компания верила в эту картину, тем сильнее перестраивала собственную стратегию и язык вокруг неё. Один бывший сотрудник вспоминал: — Я видел, как стимулы внутри Microsoft всё сильнее толкали нас к очень узкому представлению о будущем. И как технология всё больше опиралась на рассказ о ней, а не на реальность.

Сатья Наделла изменил распределение вычислительных ресурсов. GPU начали забирать у исследовательских команд Microsoft и передавать OpenAI. Все графические процессоры компании объединили в единый пул, чтобы эффективнее обслуживать генеративный ИИ. Один сотрудник Microsoft вспоминал: — Ещё в январе прошлого года обычный сотрудник Microsoft вообще понятия не имел, что такое OpenAI. А теперь сверху приходили срочные распоряжения: **найдите, как связать вашу работу с технологиями OpenAI.** Всего за несколько месяцев внутри Microsoft появилось более ста новых проектов генеративного ИИ. Люди экспериментировали с GPT-4 и ChatGPT практически во всём. Ирония была в том, что теперь сама Microsoft столкнулась со всеми проблемами, с которыми сталкивались другие компании, бросившиеся внедрять генеративный ИИ, не успев его по-настоящему понять. Команды рисков и соответствия требованиям получили массу головной боли. Некоторые сотрудники использовали не корпоративные версии моделей, а бесплатный ChatGPT от OpenAI. Тот мог использовать пользовательские данные для обучения. Возник вопрос: не утекут ли туда данные клиентов Microsoft? Не нарушит ли это требования регуляторов? Кому-то инструменты действительно сильно повышали производительность. Другим они только мешали. Один сотрудник говорил: — Возникла культура: «Используй ИИ, используй ИИ, используй ИИ». А ты сидишь и думаешь: «Это нам вообще не помогает. Мы не хотим этим пользоваться». Но оно везде. От него невозможно спрятаться.

Переход от собственных моделей Microsoft к моделям OpenAI имел ещё одно следствие. Многие сотрудники потеряли контроль над фундаментальной частью своей работы. Хотя модели OpenAI обучались и хранились на серверах Microsoft, доступ к ним внутри самой Microsoft был строго ограничен. Большинство сотрудников не могли: посмотреть данные обучения, изменить веса модели, понять внутреннее устройство системы. Им, как обычным клиентам OpenAI, выдавали доступ через API. Но в обмен Microsoft получала огромную выгоду. Azure AI резко привлекала новых клиентов. Microsoft была единственным облачным провайдером, способным совместить обычные преимущества облака — хранение, управление данными — с технологиями OpenAI. В мае 2023 года вице-президент Microsoft Эрик Бойд написал своей команде: «Azure OpenAI Service сейчас открывает нам двери ко множеству новых клиентов». В августе он снова восторженно писал: полезно иногда остановиться и осознать, как далеко они продвинулись всего за год. Число клиентов выросло **в 21 раз**. Через месяц он отметил новый рубеж. После объединения разрозненных ИИ-проектов Microsoft вокруг одной платформы и притока тысяч новых клиентов трафик вырос **в десять раз всего за девять месяцев**.

В январе 2023 года система обрабатывала около **100 миллиардов запросов в месяц**. В сентябре — **триллион.** Летом технический директор Microsoft Кевин Скотт выступил на общем собрании OpenAI. Он сказал практически без прикрас: — Мы прекратили многие собственные инвестиции в машинное обучение и ИИ настолько, что люди внутри Microsoft говорят нам: «Да пошли вы». Потому что мы забираем GPU у внутренних исследований. И добавил: — Но мы сделали эту ставку, потому что считаем, что именно вы делаете лучшую работу во всей индустрии.

ChatGPT окончательно закрепил поворот OpenAI от некоммерческой исследовательской лаборатории к коммерческому бизнесу. Альтман и другие руководители решили развивать успех. В феврале 2023 года появилась платная версия ChatGPT. В марте один за другим вышли: API ChatGPT, Whisper API, и наконец GPT-4. Один бывший сотрудник говорил: — После ChatGPT появился очевидный путь к выручке и прибыли. Уже невозможно было делать вид, что мы просто

идеалистическая исследовательская лаборатория. Перед нами были реальные клиенты, которых нужно обслуживать прямо сейчас.

Но волна продуктов снова обрушилась на команду trust and safety. Раньше OpenAI раздавала новым пользователям API бесплатные кредиты примерно на двадцать долларов. После взрывного роста ChatGPT это стало проблемой. Люди начали массово создавать новые аккаунты, чтобы снова и снова получать бонус. Другие открывали новые учётные записи, чтобы обходить блокировки. OpenAI теряла огромные деньги. Серверы становились всё дороже. А команда безопасности всё ещё насчитывала меньше двадцати человек. И вместо системной работы ей снова приходилось играть в бесконечную игру: **закрывать одну дыру — только чтобы тут же появилась следующая.**

Через несколько месяцев постоянные рывки довели руководителя команды Дэйва Уиллнера до тяжёлого выгорания. Он и несколько сотрудников ушли. К концу года команда фактически распалась. Оставшихся людей включили в более широкое подразделение безопасности под руководством давней исследовательницы OpenAI Лилиан Вэн. Некоторые бывшие сотрудники trust and safety позже считали, что многолетняя война между Applied и лагерем Safety сильно навредила их функции. Safety слишком часто связывали с «думерством» — апокалиптическими теориями о конце человечества. Из-за этого часть руководства OpenAI начала относиться с подозрением уже **ко всем разговорам о безопасности**. И без того слабая позиция trust and safety в обычных технологических компаниях внутри OpenAI стала ещё слабее.

Споры продолжались с каждым новым релизом. В том числе вокруг GPT-4. Люди из Applied считали: шесть месяцев задержки перед выпуском — это уже чрезвычайная осторожность. Некоторые сотрудники Safety отвечали: этого всё равно недостаточно для полноценного тестирования и выравнивания.

Особенно тяжёлой оставалась проблема **галлюцинаций**. Даже после целенаправленной работы с RLHF модель продолжала уверенно выдумывать факты. В ноябре 2022 года, когда пользователи ChatGPT начали обращаться к нему как к поисковой системе и заговорили о возможной угрозе Google, во внутреннем документе OpenAI отмечалось: в тестах модель галлюцинировала примерно в **30 процентах вопросов закрытого типа**. А ведь закрытый тип — это самые простые вопросы. Например:

пользователь загружает PDF и просит сделать краткое содержание; даёт набор пунктов и просит превратить их в связный текст. То есть модель должна работать только с предоставленной информацией. Это отличается от вопросов открытого типа: когда пользователь спрашивает про поп-культуру, древнюю историю, биологию — без каких-либо источников. Иными словами, даже в наиболее контролируемых условиях модель всё ещё могла выдумывать ответ примерно в трети случаев.

Тем временем GPU окончательно превратились в хроническое узкое место OpenAI. Новые исследования приходилось откладывать. Продукты — переносить. Функции — отменять. Чипов просто не хватало. После взлёта ChatGPT отраслевое издание SemiAnalysis оценивало только вычислительные расходы OpenAI примерно в **700 тысяч долларов в день**. После запуска ChatGPT Applied забрала часть GPU у Research, пообещав вернуть к определённой дате. Дата наступила. GPU не вернули. Потому что число пользователей продолжало расти. И Applied требовалось уже не меньше чипов, а ещё больше.

Давление заставило OpenAI искать более эффективные модели. Research разработала новый способ создания Transformer-моделей, которые дешевле обслуживать. Метод получил название:

DUST.

А получавшиеся модели называли по пустыням. Первой стала **Sahara** — оптимизированная версия GPT-3.5. В феврале 2023 года её выпустили под публичным названием: **GPT-3.5 Turbo.** Другая

модель получила кодовое имя **Gobi** — оптимизированная текстово-визуальная система. Третья должна была стать более эффективной версией GPT-4. Её назвали: **Arrakis** — в честь пустынной планеты из *Дюны*. Но после месяцев работы команда так и не смогла одновременно сохранить качество GPT-4 и заметно повысить эффективность. Проект съедал огромное количество вычислений. В итоге руководство его закрыло, чтобы освободить GPU для других задач.

И это стало особенно болезненным ударом в отношениях с Microsoft. Microsoft опасалась, что OpenAI может однажды уйти к другому партнёру. А OpenAI, со своей стороны, боялась: что Microsoft перестанет сотрудничать, если компания не будет достаточно стараться угодить партнёру. Arrakis стал неприятным примером. Чтобы впечатлить Microsoft, OpenAI даже изменила собственную дорожную карту. Поставила эту модель выше проектов, которые сама считала более стратегически важными. В том числе — проекта по использованию GPT-4 в поисковой системе. А в итоге Arrakis провалилась. Некоторые высокопоставленные руководители Microsoft были разочарованы.

Появилась и ещё одна неловкость. OpenAI и Microsoft всё чаще **конкурировали за одних и тех же клиентов**. После Codex и DALL-E 2 OpenAI решила сохранять право продавать свои технологии напрямую. ChatGPT и GPT-4 доказали: компания вполне способна зарабатывать сама. Но это означало: OpenAI приходит к клиенту и предлагает технологию. Та же технология передаётся Microsoft. И Microsoft приходит к тому же клиенту с почти тем же предложением.

Передача технологий Microsoft тоже становилась всё тяжелее. OpenAI ускоряла релизы. Microsoft делала то же самое. Но Microsoft была гигантом по сравнению с OpenAI. И после каждого нового продукта один сотрудник OpenAI мог получить сообщения от десятков людей из разных подразделений Microsoft. Вопросы: технические, организационные, про сроки, про интеграцию. Сотрудников OpenAI всё сильнее раздражало, что они должны поддерживать релизы Microsoft и одновременно выполнять собственную дорожную карту.

Основная работа по сглаживанию отношений легла на Миру Мурати. Она тесно взаимодействовала с Кевином Скоттом и другими руководителями Microsoft. Нужно было: согласовывать даты запусков, разводить продукты двух компаний, не допускать, чтобы они выглядели совершенно одинаковыми, наладить более разумный процесс работы. Летом 2023 года, чтобы уменьшить нагрузку от постоянных коммуникаций, команда инженеров Microsoft фактически **встроилась внутрь OpenAI**. Они получили полный доступ к необходимым системам и должны были упростить передачу технологий.

Чем глубже становилось сотрудничество, тем больше лично вовлекались Альтман и Наделла. Особенно в вопрос: **куда отдавать GPU?** Как распределять чипы? Сколько денег выделять на следующие чипы? OpenAI постоянно требовала ещё вычислений. Наделла позже рассказывал *The New York Times*, что запросы OpenAI росли так быстро, что Альтман начал звонить ему практически ежедневно: «Мне нужно ещё. Мне нужно ещё. Мне нужно ещё».

Но нужны были не только дата-центры для обслуживания ChatGPT. Для будущих моделей требовались значительно более мощные суперкомпьютеры. И две компании начали обсуждать совершенно новый проект. В OpenAI его называли: **Stargate.** В Microsoft: **Mercury.** Идея — построить один гигантский суперкомпьютер. Только строительство оценивалось примерно в: **100 миллиардов долларов.** Так империя искусственного интеллекта вновь начинала расти по тем же законам, что и древние империи. Чтобы расширяться, ей требовалось всё больше материальных ресурсов. И прежде всего — **всё больше земли.**

Глава 12

Разграбленная Земля В Сантьяго, столице Чили, после хорошего дождя смог рассеивается, и перед глазами открывается поразительный вид на Анды. Эта горная цепь тянется вдоль всей узкой полосы страны: от Патагонии на крайнем юге более чем на 4200 километров к северу — примерно вдвое дальше, чем от Майами до Бостона, — вплоть до северной границы Чили. Здесь городской пейзаж столицы постепенно уступает место бескрайней пустыне, которая, если не считать отдельных районов Антарктиды, является самым сухим местом на Земле. Горы становятся всё отчётливее, а под безжалостным солнцем их краски кажутся особенно яркими.

Задолго до того, как появилась страна под названием Чили, пустыню Атакама населяли многочисленные коренные народы. Там, где чужак увидел бы лишь суровую и бесплодную землю, они сумели устроить дом: добывали воду и минералы из глубин, выращивали урожай, разводили скот и совершали обряды предков. Затем пришли испанцы. Они расчленили регион административными границами. Старейшины коренных народов и сегодня передают устные предания о том, как жестоко подавлялось сопротивление: рассказывают, что испанцы отрезали языки тем, кто осмеливался продолжать говорить на родном языке. Именно тогда империя заложила модель отношений Чили с остальным миром, сохраняющуюся в разных формах и поныне: страна поставляет сырьё — землю, воду, энергию, минералы, — которое служит политическим и экономическим интересам других государств.

Сегодня почти 60 процентов чилийского экспорта приходится на минеральное сырьё, в основном добываемое в Атакаме. Прежде всего это медь — металл с высокой электропроводностью, необходимый практически любой электронике, — а в последние годы ещё и литий, ключевой компонент литий-ионных аккумуляторов. Именно эти ресурсы во многом приводят в движение экономику страны. В Сантьяго почти каждый знает кого-нибудь, чья жизнь подчинена ритму горнодобывающей отрасли: рабочую вахту человек проводит на севере, а в выходные возвращается в столицу.

Чили так и не удалось создать другую отрасль, которая могла бы стать сопоставимым двигателем экономики. И спустя много десятилетий после исчезновения Испанской империи Соединённые Штаты сыграли печально известную роль в сохранении такой зависимости. В 1950–1960-е годы в Чили и Уругвае набирала популярность экономика развития. Её сторонники считали, что молодой развивающейся стране для созревания необходимы жёсткое государственное регулирование и индустриализация, ориентированная прежде всего на внутренний рынок. Американским транснациональным компаниям, зарабатывавшим миллиарды на чилийских рудниках, всё больше не нравились повышавшиеся налоги и ограничения. И тогда правительство США приступило к переустройству экономической политики Чили таким образом, чтобы она лучше соответствовала интересам американского бизнеса. В 1956 году была запущена программа под названием **«Чилийский проект»**: сто чилийских студентов должны были пройти обучение в Чикагском университете под интеллектуальным руководством американского экономиста Милтона Фридмана.

Фридман был одной из крупнейших фигур экономической науки. В 1976 году он получит Нобелевскую премию, а сущность его взглядов можно было выразить названием его знаменитой статьи 1970 года в *The New York Times*: **«Социальная ответственность бизнеса состоит в увеличении его прибыли».** Фридман защищал практически всё то, чему противостояла экономика развития: минимальное государственное регулирование, максимальную свободу компаний, ориентированных на прибыль, и открытие развивающихся стран внешним рынкам, прежде всего через либерализацию экспорта. Как подробно пишет Наоми Кляйн в международном бестселлере *Доктрина шока*, вышедшем в 2007 году, «Чилийский проект» был не столько образованием, сколько идеологической обработкой. В Чикагском университете чилийских студентов — а затем и студентов из других стран Латинской Америки — специально учили критиковать экономическую политику собственных государств и видеть «фатальные недостатки» латиноамериканской модели развития.

По мере того как очередные группы выпускников, получивших прозвище **«чикагские мальчики»**, возвращались в Сантьяго, неолиберальные идеи Фридмана проникали всё глубже в интеллектуальную элиту страны. В конце концов они стали частью государственной идеологии. В 1973 году демократически избранный левый президент Чили был свергнут генералом Аугусто Пиночетом в результате военного переворота, условия для которого отчасти создавались при участии ЦРУ. Так началась почти двадцатилетняя жестокая диктатура Пиночета — и одновременно новая экономическая эпоха.

Режим поручил тем самым «чикагским мальчикам» разработать экономическую политику страны. При Пиночете в Чили приватизировали почти всё: образование, здравоохранение, пенсионную систему, даже воду. Эта стратегия действительно привела к экономическому росту. Но одновременно породила поразительное неравенство. И сегодня Чили остаётся одной из самых неравных стран мира: почти четверть национального дохода сосредоточена в руках нескольких могущественных семей, входящих в богатейший один процент населения.

Поскольку страна так и не прошла полноценной индустриализации, она осталась привязана к добывающей экономике — именно она делает Чили интересной для более могущественных игроков мировой политики. И поэтому, когда начался бум искусственного интеллекта, Чили оказалась в эпицентре **нового, ещё более масштабного этапа извлечения ресурсов**. Теперь индустрии понадобились не только медь и литий с севера страны. Ей потребовались земля, вода и энергия для всё новых дата-центров вокруг Сантьяго. В мае 2024 года правительство с гордостью объявило, что в ближайшие годы Чили примет **28 новых дата-центров** вдобавок к уже существующим 22.

Это должно было принести стране 2,6 миллиарда долларов иностранных инвестиций. Правительство представляло роль Чили как поставщика ресурсов для развития технологий в качестве национального прогресса. Но одновременно именно Чили стала одним из центров самого ожесточённого сопротивления подобной логике. По всей стране местные сообщества протестуют против того, чтобы их лишали земли, воды и других ресурсов ради проектов Глобального Севера — проектов, в которых этим людям либо вообще не находится места, либо они не получают от них практически никакой пользы.

Уличные протесты и судебные процессы уже затормозили несколько корпоративных проектов и заставили правительство обратить внимание на проблему. Более того, пример чилийских активистов вдохновляет людей в других странах. Мартин Тирони Родо, профессор Католического университета в Сантьяго и директор исследовательского центра **Futures of Artificial Intelligence Research**, изучающего ИИ с междисциплинарной и латиноамериканской точки зрения, выразил мысль, которую я снова и снова слышала, путешествуя по стране и разговаривая с местными жителями.

Главный вопрос этих движений, говорит он, заключается в следующем: **можно ли представить развитие искусственного интеллекта иначе — так, чтобы оно не строилось на непрерывном извлечении ресурсов?** — Если мы будем развивать эту технологию так же, как развивали всё прежде, — говорит Тирони, — мы опустошим Землю.

«Цифровые» технологии вовсе не существуют только в цифровом мире. И «облако» на самом деле совсем не похоже на нечто невесомое и эфемерное, как можно было бы подумать по его названию. Чтобы обучать модели искусственного интеллекта и предоставлять их пользователям, нужны вполне материальные, физические сооружения — **дата-центры**. А для обучения и эксплуатации генеративных моделей того типа, который популяризировала OpenAI, дата-центров нужно всё больше — и сами они должны становиться всё крупнее. Ещё до бума ИИ дата-центры уже стремительно росли. Когда-то они были небольшими и рассредоточенными. Их легко было спрятать внутри города: несколько полок компьютеров в подсобном помещении или несколько десятков серверных стоек в переоборудованном здании. В 2000-е годы технологические гиганты начали двигаться в совершенно другую сторону.

Они стали объединять вычислительную инфраструктуру в гигантские склады серверов, расположенные в сельской местности. Мир дата-центров раскололся на две категории: **гиперскейлеры — и все остальные.** Четыре крупнейших гиперскейлера — Google, Microsoft,

Amazon и Meta — теперь ежегодно тратят на строительство дата-центров больше, чем практически все остальные, гораздо менее известные разработчики вроде Equinix и Digital Realty, вместе взятые. Если вы никогда не видели гипермасштабный дата-центр, трудно представить его размеры. Мэл Хоган, доцент канадского Университета Куинс, изучающая искусственный интеллект, инфраструктуру и экологию, вспоминает, что около десяти лет назад, когда только начинала писать об этом, сравнивала такие центры с футбольными полями. — Теперь даже футбольных полей уже недостаточно, чтобы представить необходимый масштаб, — говорит она. Сами гиперскейлеры называют свои дата-центры **кампусами**. Это огромные участки земли, сопоставимые по площади с крупнейшими университетами Лиги плюща, на которых стоят несколько гигантских зданий, доверху заполненных рядами серверных стоек. Все эти компьютеры выделяют невероятное количество тепла — словно миллион перегревшихся ноутбуков работает одновременно. Чтобы они не расплавились от собственной температуры, здания снабжаются огромными системами охлаждения: промышленными вентиляторами, кондиционерами или установками, испаряющими воду для охлаждения серверов.

Всё вместе создаёт непрерывную какофонию гула, свиста и потрескивания. Особенно в небогатых районах этот шум слышен на километры вокруг, двадцать четыре часа в сутки, превращаясь в постоянное, изматывающее тело шумовое загрязнение. Но после появления ChatGPT понадобилось новое слово. Теперь разработчики говорят уже о **мегакампусах**. И речь идёт не столько о площади земли, сколько о чудовищном количестве энергии, необходимом для их работы. Стойка с GPU потребляет примерно втрое больше электричества, чем стойка с обычными компьютерными чипами. Причём дорого обходится не только обучение генеративных моделей. Дорого обходится и каждое их использование. По оценке Международного энергетического агентства, **один запрос к ChatGPT в среднем требует примерно в десять раз больше электроэнергии, чем обычный поиск в Google**. Ещё совсем недавно крупнейшие дата-центры проектировали на мощность около 150 мегаватт. За год такой объект мог потребить примерно столько же электричества, сколько **122 тысячи американских домохозяйств**. Теперь разработчики и энергетические компании готовятся к мегакампусам ИИ мощностью **1000–2000 мегаватт**.

Один такой объект способен ежегодно потреблять столько же энергии, сколько примерно **полтора — три с половиной Сан-Франциско**. На всей планете найдётся совсем немного мест, способных выработать и доставить такое количество электричества в одну точку. Поэтому разработчики дата-центров вместе с энергетическими компаниями по всему миру строят новые электростанции и расширяют список возможных источников энергии. После десятилетия практически неизменного спроса на электроэнергию в США аналитики Goldman Sachs назвали внезапную волну дата-центров источником **«роста потребления электричества, какого не наблюдалось целое поколение»**. Энергетические компании теперь откладывают закрытие газовых и угольных электростанций, а вместе с ним — переход к возобновляемой энергетике. Microsoft договорилась о возобновлении работы атомной электростанции Три-Майл-Айленд возле Мидлтауна, штат Пенсильвания, где в конце 1970-х произошла частичная авария реактора — крупнейшая авария в истории коммерческой атомной энергетики США. При нынешних темпах роста к 2030 году дата-центры могут потреблять **8 процентов всей электроэнергии Соединённых Штатов**, против 3 процентов в 2022 году.

А мировые вычислительные мощности ИИ могут начать потреблять больше электричества, чем **вся Индия**, третий крупнейший потребитель электроэнергии на планете. Такой масштаб — уже не просто гипермасштаб, а своего рода **мегагипермасштаб** — приводит к ошеломительным экологическим последствиям. И одновременно компании всё сильнее затемняют реальную величину этого воздействия. После работы Эммы Струбелл и того, как Гебру сослалась на неё в статье «Стохастические попугаи», технологические гиганты стали раскрывать ещё меньше технических подробностей о своих моделях. В результате оценивать и отслеживать их углеродный след стало чрезвычайно трудно. Одновременно компании усилили кампании влияния на общественность и политиков, предлагая мощные контраргументы: дата-центры станут настолько эффективными, что проблема исчезнет;

генеративный ИИ совершит революцию в борьбе с климатическим кризисом; AGI однажды решит проблему изменения климата окончательно. Последнее утверждение невозможно проверить. Но

первые два, говорит исследовательница и руководитель климатического направления компании Hugging Face Саша Луччони, вводят людей в серьёзное заблуждение. Особенно второе. Да, существуют многочисленные технологии ИИ, способные помочь экологической устойчивости, — их каталогизирует организация Climate Change AI. Но почти никогда речь не идёт о **генеративном ИИ**. — Для климата нужны модели обучения с учителем, модели обнаружения аномалий или даже статистические модели временных рядов, — говорит Луччони, одна из основательниц Climate Change AI. Все они относятся к предыдущим поколениям искусственного интеллекта — главным образом к машинному обучению. Они сравнительно малы, энергоэффективны, а некоторые могут работать даже на мощном ноутбуке. — У генеративного ИИ совершенно непропорциональный энергетический и углеродный след при очень небольшой экологической пользе, — добавляет она.

Луччони рассказывает, что её прежним коллегам, работающим теперь в закрытых корпоративных структурах, работодатели больше не разрешают писать вместе с ней статьи об экологическом воздействии ИИ. Поэтому ей приходится сотрудничать с независимыми исследователями и университетскими учёными — например, со Струбелл, чья работа когда-то вдохновила её самой заняться этой темой. В одной из статей Луччони вместе с Ясином Жернитом, руководителем направления «машинное обучение и общество» в Hugging Face, измерила углеродный след открытых генеративных моделей — чтобы хотя бы приблизительно понять, что происходит внутри закрытых компаний. Они обнаружили, что создание **тысячи текстовых ответов** генеративной модели в среднем расходует столько же энергии, сколько требуется для почти четырёх полных зарядок обычного смартфона. А создание **тысячи изображений** — столько же, сколько **242 полных зарядки смартфона**. Иными словами, одно изображение, созданное ИИ, может потребить столько энергии, сколько необходимо для зарядки телефона примерно на четверть. Исследования Луччони и Струбелл остаются одними из немногих, где вообще приводятся количественные оценки углеродного следа генеративного ИИ. И добывать такие цифры становится всё труднее. Однажды Луччони написала более чем пятистам авторам последних научных работ по машинному обучению, попросив предоставить элементарные сведения об обучении их моделей. — Я почти не получила ответов, — говорит она. — Люди либо вообще молчали, либо отвечали: это конфиденциальная информация.

При этом гиперскейлеры громко заявляют о «зелёности» своей вычислительной инфраструктуры. Но внутри Microsoft руководители признавали: переменчивая доступность солнечной и ветровой энергии плохо подходит дата-центрам, которые должны работать **двадцать четыре часа в сутки семь дней в неделю**. Непрерывность работы считается настолько критичной, что Google, Amazon, а теперь и Microsoft строят кампусы группами по три — чтобы имелся резерв для резерва на случай отказа одного объекта. Во время урагана «Ирма» во Флориде и «Харви» в Техасе, когда миллионы людей лишились электричества, некоторые больницы эвакуировали пациентов, а сотни тысяч домов и предприятий оказались повреждены или разрушены, дата-центры продолжали спокойно гудеть.

Причём настолько надёжно, что семьи сотрудников одного из них, потерявшие жильё, на время стихийного бедствия просто переселились внутрь. Но земля и энергия — лишь два компонента глобальной цепочки расширения дата-центров. Им требуются и огромные объёмы минералов, включая медь и литий, необходимые для изготовления компьютеров, кабелей, линий электропередачи, батарей и резервных генераторов. А ещё — невероятное количество **питьевой воды**. Да, именно питьевой. Её часто используют для охлаждения серверов. Вода должна быть достаточно чистой, чтобы в трубах не образовывались отложения и не размножались бактерии, а питьевая вода соответствует этим требованиям. По оценке исследователей Калифорнийского университета в Риверсайде, стремительно растущий спрос со стороны ИИ к 2027 году может потребовать во всём мире **от 1,1 до 1,7 триллиона галлонов пресной воды в год**.

Это примерно половина годового потребления воды всей Великобританией. И последствия будут распределены крайне неравномерно. Другое исследование показало, что ещё до бума генеративного ИИ пятая часть американских дата-центров уже брала воду из бассейнов, испытывающих умеренный или серьёзный дефицит из-за засухи и других причин. А в странах Глобального Юга вроде Чили основное бремя этой ускоряющейся экономики извлечения ресурсов зачастую ложится на наиболее

уязвимые сообщества. По мере того как всё больше людей видят, как дата-центры меняют их жизнь, растёт и сопротивление бесконтрольному строительству.

В ответ застройщики становятся всё изобретательнее. Они тайно входят в регионы через компании-пустышки. Жертвуют деньги местным программам, чтобы ослабить протест. Обещают властям городов экологически устойчивые объекты, а после начала строительства, когда повернуть назад становится гораздо труднее, постепенно отказываются от одного обещания за другим. В одном случае в Вирджинии жители, протестовавшие против нескольких гигантских дата-центров, обнаружили электронное письмо адвоката застройщику. Тот предлагал установить за активистами наблюдение: «Нам нужен один-два крота в этой группе».

С того самого момента, когда OpenAI сделала ставку на масштабирование, компания стремилась получить беспрецедентные вычислительные мощности. В 2019 году Брокман говорил мне: «В ИИ больше всего выигрывает тот, у кого самый большой компьютер». Вместе с Microsoft Альтман разработал план удовлетворения экспоненциально растущего аппетита OpenAI к вычислениям. Компании намеревались совместно спроектировать и построить десятки исследовательских суперкомпьютеров — по сути дата-центров, заполненных чипами Nvidia для обучения самых разных моделей OpenAI. Но главное — Microsoft должна была строить всё более крупные и мощные суперкомпьютеры для каждого следующего поколения моделей OpenAI. Альтман называл эти ступени **«фазами»**. Однажды он показал сотрудникам слайд, сравнивавший размеры каждой из них. Последняя, **Фаза 5**, выглядела головокружительно огромной по сравнению со всеми предыдущими. Суперкомпьютер, на котором обучали GPT-3, был **Фазой 1**. В нём работали десять тысяч графических процессоров V100. Объект построили в Уэст-Де-Мойне, штат Айова, куда Microsoft впервые пришла ещё в 2012 году. Компания сумела прекрасно наладить отношения с городскими властями, выплачивая, по словам тогдашнего мэра, «ошеломляющие» суммы налогов, благодаря которым город смог серьёзно модернизировать общественную инфраструктуру. За десять с лишним лет Microsoft также вложила примерно 2,5 миллиона долларов в местные общественные программы, главным образом некоммерческие. В Microsoft суперкомпьютер Фазы 1 носил кодовое имя **Odyssey — «Одиссея»**. В OpenAI его называли **Owl — «Сова»**.

У исследовательского подразделения OpenAI существовала традиция называть вычислительные кластеры в алфавитном порядке именами животных. Когда все двадцать шесть букв закончились, перешли к химическим элементам — в порядке их атомных номеров. **Фаза 2** тоже находилась в Айове. На ней обучалась GPT-4. Изначально там было восемнадцать тысяч A100 — именно это количество фигурировало в исследовательской дорожной карте 2021 года.

К завершению обучения GPT-4 их число выросло примерно до двадцати пяти тысяч. Microsoft называла Фазу 2 **Telemachus — «Телемах»**, в честь сына Одиссея. OpenAI — **Raven, «Ворон»**. **Фаза 3** переехала в Аризону. Дешёвая земля, щедрые налоговые льготы и близость к Калифорнии быстро сделали этот штат одним из любимых мест строительства дата-центров для облачных компаний. Заранее выстроив выгодные отношения с властями двух небогатых городов неподалёку от Финикса, Microsoft в 2018–2019 годах приобрела там три участка земли и, как прежде, стала жертвовать деньги местным организациям, приглушая возможные протесты населения. Совокупная площадь участков составляла почти **600 акров** — больше 450 футбольных полей. В 2024 году Microsoft добавит ещё 283 акра.

На каждом участке должен был появиться новый кампус дата-центров, обслуживающий и клиентов Azure, и OpenAI в облачном регионе **West US 3**. Фаза 3, называвшаяся в Microsoft **Inglewood**, а в OpenAI — **Whale, «Кит»**, должна была обойтись в несколько миллиардов долларов. В ней планировали разместить сотни тысяч GPU Nvidia H100, следующего поколения после A100. На этой инфраструктуре OpenAI собиралась обучать модели, которые тогда предполагала назвать GPT-4.5 и GPT-5. Фазу 4 Альтман планировал разместить в Висконсине. По его расчётам, она должна была стоить около **10 миллиардов долларов** и использовать новейшие B100 Nvidia — ещё одно поколение после H100. Сумма была невероятной: тогда даже самые дорогие гипермасштабные дата-центры обычно стоили от одного до двух миллиардов. Но Альтман и на этом останавливаться не собирался.

Он совершенно буднично рассуждал о **суперкомпьютере стоимостью 100 миллиардов долларов** — Фазе 5. Однако после ошеломляющего успеха ChatGPT пришлось умерить аппетиты. Огромное количество чипов требовалось просто для обслуживания пользователей ChatGPT, и Microsoft не успевала покупать их достаточно быстро даже для завершения Фазы 3. «Кит» пришлось разделить на три отдельных кластера: **Beluga — «Белуха», Narwhal — «Нарвал» и Orca — «Косатка».** Полностью достроить их планировали в течение 2024 года. Никто ни в Microsoft, ни в OpenAI даже не знал, возможна ли Фаза 5 технически. Как позднее сообщало издание *The Information*, по проектам Microsoft и OpenAI объект стоимостью сто миллиардов долларов мог потребовать до **5000 мегаватт мощности**. Почти столько же в среднем потребляет весь Нью-Йорк. И хотя с точки зрения обычной экономики такой проект выглядел сомнительно, главным ограничением были вовсе не деньги.

Главным ограничением стала **энергия**. — У нас заканчиваются земля и электричество, — говорил один сотрудник OpenAI. В обеих компаниях понимали: Фаза 5 станет возможной только после какого-то технологического прорыва. Либо Microsoft и OpenAI придётся разбить суперкомпьютер на несколько кампусов и научиться обучать одну модель одновременно в удалённых друг от друга местах.

Либо, как любил иногда говорить Альтман, проблема однажды сама исчезнет благодаря прорыву в **термоядерной энергетике**. Он временами делился с сотрудниками OpenAI оптимистичными новостями о Helion Energy — стартапе термоядерной энергетики, который являлся крупнейшей личной инвестицией Альтмана. Microsoft уже обязалась покупать у Helion электричество, когда заработает её первая станция мощностью 50 мегаватт. Позже *The Wall Street Journal* сообщит, что OpenAI и Helion также обсуждали отдельную сделку, хотя Альтман самоустранился от переговоров из-за конфликта интересов. Но Альтман и другие руководители практически никогда не говорили на общих собраниях OpenAI об экологической цене этих дата-центров. Когда GPT-4 обучалась в Айове, штат уже второй год страдал от засухи.

Позднее Associated Press сообщило, что только за один месяц обучения модели дата-центры Microsoft использовали около **11,5 миллиона галлонов воды** — шесть процентов всего водопотребления округа. А GPT-4 обучалась там три месяца. Представитель Microsoft заявил, что компания работает над повышением водной эффективности на 40 процентов по сравнению с уровнем 2022 года и намерена к 2030 году в глобальном масштабе восстанавливать больше воды, чем потребляет, уделяя особое внимание регионам с дефицитом водных ресурсов. Аризона тем временем переживает собственный тяжёлый водный кризис. В 2022 году, когда Microsoft закладывала основу Фазы 3, исследование в *Nature Climate Change* показало, что юго-запад США столкнулся с самой тяжёлой засухой более чем за **тысячу лет**. В сочетании с серьёзными ошибками управления она довела реку Колорадо, от которой зависят Аризона и ещё шесть штатов, до опасно низкого уровня. Без радикальных мер река однажды может просто перестать течь.

Нехватка воды усиливает и энергетический кризис. Изменение климата приносит региону всё новые температурные рекорды, семьи всё интенсивнее пользуются кондиционерами. А часть электроэнергии производится гидроэлектростанциями, включая плотину Гувера, и охлаждаемыми водой атомными электростанциями. То есть региону требуется **вода, чтобы производить больше энергии**. В 2023 году агломерация Финикса установила сразу несколько рекордов жары.

Это оказался и худший год по числу смертей, связанных с жарой: их количество выросло как минимум на 30 процентов по сравнению с 2022 годом и превысило шестьсот. Том Бушацке, директор Управления водных ресурсов Аризоны, сформулировал ситуацию коротко: — Всё сходится в крайне неблагоприятном направлении. Но Альтмана больше беспокоило другое. **Он терял терпение.** В марте 2024 года Meta, проспав первые годы гонки генеративного ИИ, внезапно вышла на поле с чрезвычайно агрессивными планами. Компания собиралась запустить **350 тысяч H100** в рамках ещё более грандиозного расширения инфраструктуры. Это было больше GPU, чем имелось в распоряжении OpenAI. Альтман был недоволен. Ему казалось, что Microsoft движется слишком медленно. И из-за этого OpenAI теряет своё конкурентное преимущество.

Когда Сония Рамос была ребёнком, она стала свидетелем трагедии, которая определила всю её дальнейшую жизнь. Она родилась в семье шахтёров на севере Чили. Её отец работал на американскую медную компанию, и Сония росла среди детей других рабочих. В 1957 году часть рудника Чукикамата обрушилась. Несколько человек погибли. Десятки получили ранения. Её отец выжил. Но Сония помнит, как после аварии вокруг неё развернулась настоящая человеческая катастрофа. Пострадавшие семьи стремительно скатывались в крайнюю нищету. Дети худели от голода. Четыре десятилетия спустя, когда Рамос станет одним из самых активных и громких голосов коренных народов Чили, протестующих против социального, культурного и экологического разрушения, которое несёт добыча полезных ископаемых, она будет вспоминать именно тот урок. Горнодобывающая промышленность — это система. И если оставить её без контроля, она будет искать прибыль **любой ценой**. — Рабочего для этой системы не существует, — говорит она.

Никому из погибших не устроили достойной церемонии памяти. Ни одна из семей не получила компенсации. — Там нет человечности. Чили — крупнейший производитель меди в мире, обеспечивающий примерно четверть мировых поставок. С самого начала медные рудники меняли не только землю. Они меняли и общества, зависящие от этой земли. Иногда последствия видны буквально. Чукикамата сегодня — крупнейший в мире открытый медный карьер. Гигантская рана в земной поверхности, которую взрывы регулярно делают всё глубже. Вывороченная порода складывается в огромные горы, словно памятники пустотам, из которых её извлекли. Эти отвалы постепенно погребают остатки города, покинутого жителями после того, как медный карьер начал буквально его поглощать. Добыча меди высушила и региональные водные ресурсы. Когда-то одна иностранная транснациональная корпорация использовала столько воды, что полностью истощила целый бассейн на соседнем солончаке — *саларе*, уничтожив богатую местную экосистему. Менее заметны следы мышьяка, которые горнодобывающая промышленность оставляет в воздухе и воде. Их связывают с повышенной заболеваемостью раком по всему северу страны. Не менее разрушительно добыча полезных ископаемых изменила жизнь коренных народов и посеяла раздоры между общинами.

Атакаменьо — общее название различных коренных народов этого региона — больше не могут обеспечивать себя земледелием и скотоводством. Их земли лишились воды и минералов. Города и поселения погрузились в глубокую нищету. Растут преступность, депрессия, алкоголизм и правонарушения. Людям не хватает еды, водопровода, нормального здравоохранения и образовательных возможностей. При этом они почти ничего не получили от миллиардов долларов прибыли, которые их земля принесла кому-то другому. Многие теперь вынуждены работать на ту самую индустрию, которая отняла у них территории, и лечиться в маленьких медицинских пунктах, которые эта индустрия финансирует. Если прежде общины были более едины, теперь они всё чаще спорят между собой из-за сокращающихся ресурсов. Литий здесь обнаружили сравнительно недавно. В 1960-е годы американская компания искала воду, необходимую для добычи меди, и пробурила скважины в солончаках. Под поверхностью обнаружились огромные концентрации лития, растворённого в маслянистом рассоле. Так открылся новый фронт добычи. И ещё быстрее началось разрушение экосистем. Сегодня Чили производит примерно треть мирового лития, уступая только Австралии.

Основная добыча идёт на **Салар-де-Атакама**, крупнейшем солончаке страны. Рассол выкачивают на поверхность и разливают по сверкающим бирюзовым бассейнам, где солнце постепенно испаряет воду, оставляя литий и другие вещества в кристаллической форме. Когда-то над этими солончаками собирались стаи розовых фламинго. Для атакаменьо они были почти духовными братьями. Теперь фламинго исчезли. Маленькая дочь одного из лидеров общины Пейне знает о них только из рассказов предков. И ещё у неё есть плюшевый фламинго. Годами атакаменьо слышали разные объяснения, которыми оправдывалась добыча. В 2022 году, когда Евросоюз принял новые программы энергетического перехода и спрос на литий резко взлетел, компании и политики как в Чили, так и за рубежом начали прославлять горнодобывающую отрасль страны как ключ к **лучшему будущему**.

Коренные жители смотрели, как их земля и общины распадаются, и задавали простой вопрос: **Лучшее будущее — для кого?** — У местных жителей никогда нет возможности самим решать свою

судьбу независимо от сил экономики и международной политики, — говорит микробиолог Кристина Дорадор, живущая на севере Чили и изучающая его уникальное биоразнообразие. Теперь те же аргументы повторяют применительно к генеративному ИИ. Ускоренная добыча меди и лития, необходимая для строительства мегакампусов, электростанций и тысяч километров новых линий электропередачи, в версии Кремниевой долины тоже ведёт человечество к светлому будущему. Следовательно, препятствовать добыче — значит препятствовать прогрессу всего человечества.

Но коренные жители, подчёркивает Рамос, выступают вовсе не против добычи как таковой. — Наши предки сами были горняками. Именно они когда-то открыли здесь медь. Проблема — **в масштабе**. Масштаб поглотил всё. Он поставил север и всю страну в полную зависимость от добывающей отрасли и не дал возникнуть альтернативной экономике. Он задушил способность Чили — и остального мира — даже представить себе иной путь развития, при котором прогресс возможен без разграбления природных ресурсов. И тот же масштаб позволяет производить гигантские генеративные модели ИИ, которые, в свою очередь, воспроизводят расистские стереотипы о коренных народах — тех самых людях, которые уже страдают от физической инфраструктуры, необходимой для создания этих технологий. На выставке 2023 года в Бразилии, организованной в том числе одним из чилийских университетов, был показан огромный разрыв между реальной сложностью и богатством культур коренных народов Латинской Америки и жалкими образами, которые генерировали Midjourney и Stable Diffusion.

ИИ изображал их примитивными, технологически беспомощными людьми. В последние годы сопротивление атакаменьо усиливается. На домах появляются чёрные флаги — символ протеста против эксплуатации земли и общин. Активисты перекрывают дороги, по которым автобусы и грузовики компаний едут к шахтам. Они нанимают адвокатов и добиваются соблюдения своих прав коренных народов, закреплённых международным правом, включая культурный и территориальный суверенитет. Компании и правительство Чили теперь вынуждены приглашать их за стол переговоров.

Одно из главных требований — провести независимые исследования состояния экосистем Атакамы, подсчитать потерю воды и определить, какой ущерб уже стал необратимым. У Рамос есть собственный фонд. Его задача, по её словам, — соединить **«наследственное и современное»**, поддерживая научные исследования природного богатства пустыни Атакама. Из-за экстремальных условий здесь существуют уникальные микробные сообщества, которых нет больше нигде на Земле. Они потенциально могут оказаться полезными для новых лекарств и источников энергии.

По тем же причинам Атакаму десятилетиями изучают как земной аналог марсианского климата. Рамос надеется, что открытия помогут доказать ценность сохранения её прекрасной родины. В противоположность идеологии сверхбыстрого «прогресса», оправдывающей извлечение ресурсов, она ищет иное понимание прогресса — основанное на **исцелении, устойчивости и восстановлении**.

Пока Рамос продолжает борьбу на севере, в самом сердце Чили идёт другое сражение — уже непосредственно вокруг дата-центров, которые правительство так охотно принимает. Чем быстрее гиперскейлеры растут и чем сильнее им начинает не хватать земли и электроэнергии в привычных регионах, тем агрессивнее они претендуют на ресурсы новых территорий по всему миру. Только Microsoft в 2024 финансовом году потратила более **55 миллиардов долларов**, почти четверть заявленной выручки, на то, что аналитическая компания SemiAnalysis назвала: «крупнейшим строительством инфраструктуры, которое когда-либо предпринимало человечество».

Google тем временем в третьем квартальном отчёте 2024 года заявила, что собирается увеличить расходы на дата-центры примерно до **50 миллиардов долларов** за финансовый год. Meta сообщила, что её расходы на дата-центры и инфраструктуру могут достичь **40 миллиардов** и в следующем году вырастут ещё больше. Редким туманным днём в июне 2024 года Александра Арансибия ведёт нашу машину через Киликуру — муниципалитет на окраине Сантьяго, где она живёт и работает депутатом местного совета.

Она хочет показать мне то, что считает самым наглядным символом колоссального неравенства сил между американскими технологическими гигантами и своей общиной. Менее чем в получасе езды отсюда находятся самые живописные кварталы Сантьяго — с кафе в европейском стиле и веганскими

ресторанами. Но здесь дороги разрушаются из-за отсутствия ремонта. По обочинам лежат горы мусора на нелегальных свалках, которые контролирует местная мафия. Мы проезжаем кладбище домашних животных и останавливаемся возле участка, похожего на заброшенный пустырь. Кочки травы. Редкие кусты.

Несколько измученных нехваткой питательных веществ деревьев торчат из земли. Обычно почва здесь настолько сухая, что напоминает Атакаму. Сегодня дождь превратил её в грязь. Посреди участка стоит фиолетовый щит: **«Добро пожаловать в городской лес Киликура».** Здесь объясняется, что проект был начат Google в 2019 году в качестве благодарности общине за размещение дата-центра. На плакате даже нарисована схема его «преимуществ». Слева Киликура изображена как промышленная зона, забитая заводами, выбрасывающими парниковые газы и загрязняющими воздух. Справа — цветущий лес под благодатным дождём из большой тучи с надписью:

SMOG.

Google хвастается этим лесом на своём сайте и в пресс-релизах. Когда я попросила представителя компании в Чили дать интервью о деятельности Google в стране, вместо этого она прислала мне тщательно подготовленные информационные материалы.

Позднее добавила, что для каждого нового дата-центра Google создаёт специальную программу помощи местным сообществам — поддерживает образование, экологию, доступ к интернету и здравоохранение. В Киликуру компания, по её словам, вложила более 1,2 миллиона долларов. В материалах о лесе говорится, что жители используют это зелёное пространство. Но здесь **нет жителей**. Место находится слишком далеко от автобусных маршрутов. Вокруг практически нет жилых домов. За пределами скромного участка — настолько маленького, что на нём не поместился бы сам дата-центр Google, — бродит с десяток бездомных собак, лают и роются в мусоре. Представитель Google сказала, что лес сейчас модернизируется, чтобы «развивать опыт взаимодействия с сообществом». Арансибия смотрит вокруг и криво улыбается: — Ну что, чувствуете себя как в Кремниевой долине?

Арансибия только поступила в университет, когда впервые осознала: **Киликура — это место, куда всё выбрасывают.** Каждый день она ехала на занятия через районы муниципалитета, о существовании которых раньше почти не знала. Повсюду лежали груды мусора. До этого она никогда по-настоящему не воспринимала Киликуру как «дом». Просто место, где жила её семья, — бедный, ничем не примечательный рабочий пригород, куда родители переехали, когда она была маленькой. Но вид родного района, превращённого буквально в свалку, неожиданно пробудил в ней желание вернуть этой земле прежнюю красоту. Всего два десятилетия назад, когда Арансибия была ребёнком, Киликура оставалась в основном сельской местностью: зелёные пастбища, сверкающие водно-болотные угодья, множество птиц, животных и растений. Биоразнообразие здесь отличалось от Атакамы, но было не менее богатым. Потом пришла мусорная мафия. За деньги она разрешала кому угодно из Сантьяго свозить отходы в Киликуру.

Некоторые свалки существуют настолько давно, что уже заросли травой и сорняками и выглядят как странные, деформированные холмы, постепенно поглощающие пейзаж. Затем появились промышленные предприятия — в том числе пивоваренные компании и застройщики. Они забрали ещё больше земли и начали выкачивать воду из болот. Сегодня лишь небольшие островки зелени, разбросанные по территории муниципалитета площадью около 57 квадратных километров, позволяют представить, какой Киликура была когда-то. Именно в этот район в 2012 году пришла Google, чтобы построить свой первый дата-центр в Латинской Америке. На сайте Google объект, введённый в эксплуатацию в январе 2015 года, с гордостью представляется одним из самых эффективных и экологически чистых дата-центров континента.

В те годы почти никто в Киликуре не обращал на него внимания. А уж в богатых кварталах Сантьяго никто тем более не интересовался происходящим в Киликуре. Местные жители, каждый день проезжавшие мимо объекта на автобусе по дороге на работу, думали, что это обычная фабрика — возможно, там делают пиво или продукты и создают рабочие места. На самом деле дата-центр Google — как и большинство подобных объектов — почти не создавал постоянных рабочих мест после

окончания строительства. В 2024 году одна из немногочисленных вакансий для механического техника была опубликована на сайте Google **только на английском языке**.

В объявлении прямо говорилось: резюме на любом другом языке рассматриваться не будут. Активисты отмечают, что и других преимуществ для населения дата-центр почти не принёс. Рядом государственным школам по-прежнему не хватает нормального интернета и устройств, с помощью которых дети могли бы им пользоваться. Зато появление Google сделало Киликуру и вообще Сантьяго привлекательной точкой физической экспансии Кремниевой долины. В 2019 году Google объявила о строительстве второго латиноамериканского дата-центра в столичном регионе. Вскоре Microsoft и Amazon сообщили, что тоже приходят в Чили. Правительство радостно приветствовало их, рекламируя страну как безопасное и стабильное место для прямых иностранных инвестиций в Латинской Америке — регионе, которому часто приписывают нестабильность правительств и социально-экономические потрясения. В 2020 году правительство пошло ещё дальше.

Оно объявило о строительстве нового подводного кабеля — своего рода **магистрали данных**, которая должна напрямую соединить Азиатско-Тихоокеанский регион с Америкой через центральное побережье Чили неподалёку от Сантьяго. Чили собиралась стать мировым центром цифровой инфраструктуры. Google поддержала проект. Но в июле 2019 года, когда Google начала оформлять документы на строительство второго дата-центра, за процессом уже внимательно следила группа местных жителей. Для нового объекта компания выбрала **Серрильос**, ещё один рабочий муниципалитет на границе Сантьяго. Как и Киликура, Серрильос десятилетиями оставался забытым и заброшенным. С 1930-х по 1990-е здесь работал цементный завод бельгийской компании, загрязнявший район смертельно опасным асбестом. Один чилийский историк назвал произошедшее: «крупнейшим промышленным геноцидом в истории страны».

До сих пор жители Серрильоса умирают от рака чаще среднего. Но у Серрильоса есть и уникальное преимущество. В стране, где вода приватизирована, именно здесь действует **единственная в Чили государственная система водоснабжения**. Она использует местные подземные воды, обеспечивая соседние районы, а в экстренных случаях и другие части страны. Это сочетание — многолетнее пренебрежение со стороны властей и наличие драгоценного источника воды — создало идеальную среду для появления экологических движений. Активисты привыкли следить за происходящим и яростно защищали местные ресурсы. Тем летом Google передала в чилийское экологическое ведомство документы для утверждения нового дата-центра. Обычно такое согласование было почти формальностью. Активистская группа **MOSACAT**, защищавшая водные ресурсы, начала читать все **347 страниц** заявки. Где-то глубоко внутри документа обнаружилась цифра: Google планировала расходовать на охлаждение серверов примерно **169 литров чистой питьевой воды в секунду**.

Это означало, что за год дата-центр мог потребить **более чем в тысячу раз больше воды, чем всё население Серрильоса — около 88 тысяч человек**. Для MOSACAT это было неприемлемо. Тем более что воду собирались брать непосредственно из государственной системы водоснабжения Серрильоса. И происходило это в тот самый момент, когда всё питьевое водоснабжение страны оказалось под угрозой. В 2019 году Чили — как Айова и Аризона — уже девятый год подряд переживала разрушительную и исторически беспрецедентную **мегазасуху**.

Тания Родригес, участница MOSACAT, кладёт мне на колени все 347 страниц экологической документации Google — распечатанные, сброшюрованные спиралью между двумя синими пластиковыми обложками. Тяжёлый том с глухим стуком падает мне на ноги. Почти физическое воплощение того, как Кремниевая долина использует технические знания, чтобы оправдывать централизованное принятие решений. Снизу торчат аккуратно подписанные стикеры. На одном написано: **Agua potable — питьевая вода.** Он отмечает страницы, где говорится о необходимости использовать пресную воду для охлаждения дата-центра. MOSACAT была основана в 2019 году. Активисты из разных движений — женского, жилищного, рабочего, экологического — объединились в единый коллектив.

Многие познакомились во время протеста против нелегального горнодобывающего проекта. По словам группы, тогда им удалось выгнать горняков, закрыть проект и добиться присвоения территории статуса природного заповедника. Вскоре после этого знакомый активистов, который теперь заседает в национальном конгрессе Чили, предупредил их о проекте дата-центра Google и посоветовал обратить внимание на объём планируемого водопотребления. Участники MOSACAT не были специалистами по технологиям. Но они прочитали каждую страницу. Разбирались в сложных диаграммах. Изучали незнакомую техническую терминологию. Делали подробные заметки.

Выучили устройство дата-центров и систем охлаждения. Им нужно было подготовиться к противостоянию с Google. Когда я спрашиваю Родригес, как им вообще удалось разобраться во всём этом, она весело смеётся. — Пришлось всем вместе. В группе чуть больше десятка добровольцев. Всю эту работу они выполняли в свободные часы — между основной работой и семейными обязанностями. Обычно проекты, связанные с использованием воды, получают согласование в Чили медленно. Google одобрили быстро. Даже несмотря на череду чрезвычайных ситуаций, связанных с засухой.

Сначала MOSACAT попыталась спорить с местным партнёром Google — чилийской инвестиционно-сервисной компанией Dataluna. Первая встреча, по словам активистов, прошла ужасно. Представители Dataluna, казалось, сами плохо понимали проект и отрицали, что дата-центр вообще будет пользоваться пресной водой. Тогда MOSACAT обратилась к местным властям. Мэр и городской совет тоже встречались с Dataluna — и, как вспоминают активисты, также ошибочно считали, что для охлаждения будет использоваться только сточная вода.

После объяснений MOSACAT власти встревожились и потребовали от Dataluna разъяснений. Вопрос постепенно поднялся от Dataluna к чилийскому подразделению Google, а затем дошёл до штаб-квартиры Google в Маунтин-Вью, Калифорния. В октябре 2019 года Google направила в Серрильос двух инженеров и адвоката для встречи с местными жителями. В день их приезда MOSACAT развесила протестные плакаты по всему маршруту, которым представители Google должны были ехать к месту встречи. И на саму встречу активисты пришли не одни. Среди примерно двух десятков жителей присутствовали представители ещё шести общественных и активистских организаций. Инженеры Google были, как вспоминают в MOSACAT, высокими американцами — *gringos* — и говорили только по-английски. Они сделали чрезвычайно техническую презентацию. Адвокат Google одновременно выполнял роль переводчика. Участники MOSACAT, знавшие английский, утверждают, что слышали, как адвокат неправильно переводил их слова. В какой-то момент представители Google попытались успокоить общину, предложив посадить здесь **городской лес**, такой же, какой компания уже «подарила» Киликуре. Активистов это совершенно не впечатлило. Для них было очевидно: Google пришла не для того, чтобы действительно слушать людей и обсуждать, чего они хотят. — Они приехали нас запугать, — говорит Алехандра Салинас, участница MOSACAT и депутат совета соседнего муниципалитета Майпу. — Только подумайте: они предлагают нам деревья и одновременно высушивают нашу землю.

Жителям Серрильоса не нужны были деревья. Им не нужно было, чтобы Google строила им парк — словно столичный регион Сантьяго настолько отсталый, что парков там никогда не видели. Им нужно было, чтобы Google перестала относиться к их земле как к месту, откуда можно выкачать драгоценную воду и другие ресурсы. Чтобы компания перестала воспринимать местное население как случайных наблюдателей — и признала его полноценным участником решений о проектах на собственной территории. — Мы знаем, что кормим мир, что поставляем сырьё вроде меди и лития, — говорит Салинас. — Никто не говорит: это наше сокровище, мы никому ничего не дадим. Конечно, мы можем помогать друг другу. Но нельзя прийти, забрать воду — то, без чего невозможна жизнь, — и оставить нас ни с чем.

В то время Чили переживала многомесячный политический взрыв, получивший название **Estallido Social — «Социальный взрыв»**. Массовые, временами жестокие протесты начались именно в октябре 2019 года и проходили почти каждую неделю. Это был коллективный протест против безработицы, приватизации и глубочайшего неравенства. Тысячи людей получили ранения. Десятки погибли. В конце 2019 года по всей стране начали проходить местные референдумы, призванные определить направления политических реформ. MOSACAT воспользовалась референдумом в

Серрильосе, добавив туда вопрос: **согласны ли жители на строительство Google объекта, который будет расходовать столь огромное количество их воды?** Активисты развернули кампанию.

Раздавали листовки на перекрёстках. Стучали в двери. Развешивали плакаты по всему району. В декабре 2019 года MOSACAT победила. Предложение о строительстве дата-центра было отвергнуто небольшим большинством голосов.

На следующий год муниципальные власти вместе с MOSACAT подали иск против проекта в экологический суд. На той встрече Google в конце концов смягчила тон. По словам MOSACAT, представители компании стали гораздо дружелюбнее и проявили больше готовности к переговорам, когда поняли, что некоторые жители прекрасно понимают английский. Но, как говорили мне профессор Йельского университета Хулиан Посада и адвокат Мерси Мутеми, представлявшая интересы Мофата Окиньи, у стран Глобального Юга всегда остаётся опасение:

если слишком сильно сопротивляться компании из Кремниевой долины, она просто соберёт вещи и унесёт свои деньги куда-нибудь ещё. И именно это в итоге произошло. Пока проект всё сильнее буксовал, а потребность Google в вычислительных мощностях росла, компания объявила, что следующий дата-центр в Латинской Америке построит уже не в Чили. А в другой стране.

В Уругвае — небольшой стране с населением 3,4 миллиона человек — государственная телекоммуникационная компания Antel располагает тремя дата-центрами, обеспечивающими интернетом и мобильной связью всю страну. Вместе они занимают всего около **пяти тысяч квадратных метров**. Один из них — площадью менее тысячи квадратных метров — находится прямо на обычной жилой улице Монтевидео. Он выше и шире соседних домов, но всё же органично вписан в район. Дата-центр питается от двух линий электропередачи, отделённых от жилой сети. Только 30 процентов площади занимает вычислительная техника. Остальные 70 процентов — административные помещения, электрощитовые и инженерные комнаты. Компьютеры нагреваются, но не настолько сильно, чтобы для охлаждения потребовалась вода. Хватает воздуха. Они издают тихий гул, почти незаметный в дневном городском шуме. Ночью он становится достаточно раздражающим, чтобы две соседние семьи время от времени приходили жаловаться. Управляющий дата-центром Хавьер Эчеверрия выглядит немного смущённым, когда рассказывает об этом.

Он уже работает с местными исследователями над снижением шума и модифицировал систему охлаждения, чтобы она работала тише. Такая готовность отвечать на жалобы разительно отличается от того, через что пришлось пройти MOSACAT, чтобы просто заставить американскую корпорацию обратить на себя внимание. В получасе езды, сразу за городской чертой, правительство выделило огромную территорию под научный парк **Parque de las Ciencias**. Он действует как *zona franca* — свободная экономическая зона. Некоторые с горечью называют его **zona America**, потому что большинство расположенных там компаний американские и не платят государству налогов. Сам парк тоже немного напоминает Америку: ухоженные зелёные газоны, симметричная планировка, величественный фонтан в форме солнечных часов — всё напоминает торжественную эстетику Национальной аллеи в Вашингтоне. На главной странице сайта парк рекламирует Уругвай как политически и экономически стабильную страну, располагающую большим количеством земли, энергии и воды. И вот в 2021 году, когда успех GPT-3 разогрел интерес к резкому масштабированию ИИ, Google купила здесь **29 гектаров земли**.

Это в **58 раз больше**, чем вся площадь трёх дата-центров Antel вместе взятых. Здесь должен был появиться новый дом для второго латиноамериканского дата-центра Google. Но в действительности воды в Уругвае тогда вовсе не было в изобилии. Как Чили. Как Айова. Как Аризона. Страна тоже переживала разрушительную засуху. Дефицит воды стал настолько тяжёлым, что фермеры теряли целые урожаи. Сельскохозяйственный ущерб превысил миллиард долларов. К лету 2023 года власти Монтевидео начали смешивать обычную питьевую воду с загрязнённой солёной водой. Из кранов в домах потекла мерзкая коричневатая жидкость с резким химическим запахом. Кто мог себе позволить — покупал бутылочную воду. Мылись при открытых окнах, стараясь не вдыхать слишком много потенциально канцерогенных испарений. Кто не мог себе этого позволить, продолжал пить воду из-под крана.

У людей появлялись боли в животе, кожные высыпания, обострялись хронические заболевания. Росло и число выкидышей. А после тяжелейшей пандемии огромное количество людей не могли позволить себе покупать воду. В то самое время, когда Кремниевая долина переживала небывалый взлёт и рыночная стоимость Google и Microsoft превышала два триллиона долларов — во многом благодаря переходу компаний на удалённую работу и облачные сервисы, — число нелегальных поселений в Уругвае резко выросло. В местных бесплатных столовых, *ollas*, детям иногда приходилось отказывать в еде. Те, кто управлял этими столовыми, сами жили в нищете и едва сводили концы с концами. Фабиана, шумная и энергичная руководительница одной *olla*, живущая в нелегальном поселении и с любовью прозванная **Reina Madre — Королева-мать**, неожиданно замолкает, вспоминая эти годы. — Говорить человеку: «У меня нет даже маленькой тарелки еды для твоего ребёнка…» Она не заканчивает предложение. — Это было ужасно. Фабиана всю жизнь жила в бедности. В семь лет ей приходилось подметать полы в борделе, чтобы выжить.

Но годы пандемии и засухи она всё равно считает одними из самых страшных в своей жизни. Водный кризис возник из сочетания изменения климата и провальной системы распределения пресной воды. В Уругвае более **80 процентов пресной воды расходует промышленность**, а не население. Особенно много идёт на товарное сельское хозяйство: промышленные поля сои и риса, плантации деревьев для производства бумаги. Большинство таких хозяйств принадлежит не местным, а транснациональным компаниям. Они экспортируют выращенную продукцию и практически не несут ответственности за состояние природы Уругвая. Их деятельность истощает почвы, затрудняя выращивание обычной пищи, и загрязняет водоёмы огромными объёмами удобрений. Уругвай стал одним из мировых лидеров по потреблению удобрений на душу населения.

И одновременно здесь наблюдаются необычно высокие показатели раковых заболеваний. Социолог Даниэль Пенья из Республиканского университета Монтевидео, много лет изучающий политические последствия такого экологического экстрактивизма, напрямую связывает происходящее с колониальной историей Уругвая. Он ездит по стране на стареньком пикапе, разговаривает с фермерами и жителями беднейших районов, документируя, как промышленность вытесняет их на обочину. Как и в Чили, иностранные транснациональные корпорации по-прежнему стоят выше местного населения в политической иерархии. Во время засухи промышленность продолжала беспрепятственно забирать воду из главной реки — Рио-Санта-Лусия, той самой, которая питает систему водоснабжения Монтевидео. Двадцать лет назад, после мощной экологической кампании, Уругвай стал **первой страной мира, закрепившей право на воду в конституции как право человека**. Теперь возникла горькая ирония. Во время дефицита сильнее всего ограничили именно питьевую воду для людей.

А промышленность продолжала получать её почти без ограничений. Поэтому, когда появилась Google, Пенья насторожился. Регулярно просматривая сайт Министерства окружающей среды, где публиковались крупные промышленные проекты, он обнаружил заявку на строительство дата-центра. Пенья уже читал о том, как гиперскейлеры используют питьевую воду даже во время тяжёлых засух, и знал о сопротивлении таких групп, как MOSACAT. Но когда он загрузил документы Google, показатели водопотребления оказались помечены: **конфиденциально.** Пенья подал официальный запрос на доступ к общественной информации. До этого он успешно делал это раз двадцать. Но министерство и теперь отказалось раскрывать цифры, заявив, что они являются коммерческой тайной. Пенья задумался: **что они скрывают?** И какой прецедент это создаст для остальных облачных компаний, которые неизбежно последуют за Google в Уругвай? Тогда он вспомнил конституционную норму о праве на воду. Знакомый адвокат согласился бесплатно вести дело.

И Пенья подал на министерство в суд. В марте 2023 года, спустя четыре месяца, он неожиданно выиграл. Министерству пришлось раскрыть данные. Дата-центр Google планировал потреблять **два миллиона галлонов воды в сутки непосредственно из питьевого водоснабжения**. Столько ежедневно расходуют примерно **55 тысяч человек**. И вскоре после этого из кранов значительной части Монтевидео потекла солёная вода. Результат оказался взрывоопасным. Тысячи уругвайцев вышли на улицы, протестуя против Google и других отраслей, ради которых правительство растратило драгоценные пресноводные ресурсы. Когда я приехала в страну в июне 2024 года, лозунги тех

протестов всё ещё можно было увидеть на стенах и придорожных ограждениях. На одном было написано: **«Это не засуха. Это грабёж».**

Пенья смотрит на дата-центры примерно так же, как Рамос смотрит на шахты. Проблема не в самой инфраструктуре. Проблема — **в масштабе**, в котором Кремниевая долина пытается её строить. Именно этот масштаб заставляет Google и Microsoft расширяться в Чили и Уругвае даже тогда, когда сами страны переживают тяжелейший дефицит ресурсов. Именно масштаб заставляет Google занимать в 58 раз больше земли, чем Antel. И именно он позволяет корпорациям почти не отчитываться перед местными жителями. — Это экстрактивистские проекты, — говорит Пенья. — Они приходят на Глобальный Юг за дешёвой водой, землёй без налогов и очень низкооплачиваемым трудом. А потом почти ничего не дают нашей стране. Они даже не улучшают нам доступ к интернету.

В конце 2023 года Google тихо изменила проект. Теперь дата-центр в Уругвае должен был использовать **систему охлаждения без воды**, а его размер уменьшался до трети первоначального. Но Пенья считает, что борьба не закончена. Теперь правительство отказывается раскрывать прогнозируемое энергопотребление нового проекта — опять же под предлогом коммерческой тайны. Пенья также направил министерству петицию с более чем четырьмястами подписями. Он требует полноценной экологической и социальной экспертизы **всей цепочки поставок** дата-центра: где добываются минералы для его оборудования; в каких условиях трудятся рабочие; сколько углерода будет выброшено; куда отправят электронные отходы; как предотвратят попадание токсичных веществ в чьё-нибудь местное сообщество. Большинство этих последствий ощутит не сам Уругвай. Но Пенья чувствует солидарность со странами, на которые они лягут. — Как правило, все они находятся на Глобальном Юге, — говорит он. — Мы все в итоге получаем одни и те же последствия, просто на разных звеньях глобальной цепочки поставок.

В 2024 году экологический суд Чили постановил: Google **не может строить в Сантьяго дата-центр с водяным охлаждением**. Представитель Google Chile заявил, что компания сохраняет приверженность Чили и Латинской Америке и, когда возникнет необходимость, подаст документы на строительство в Серрильосе объекта с воздушным охлаждением. Но других гиперскейлеров это не остановило. В 2022 году Microsoft окончательно выбрала место для своего чилийского дата-центра — вскоре после второй инвестиции в OpenAI. И где же? Снова в родном городе Арансибии. **В Киликуре.**

В мире дата-центров Microsoft сравнительно поздно включилась в гонку. До инвестиций в OpenAI Google заметно опережала её по количеству строящихся объектов по всему миру. Но внезапный взрыв спроса на вычислительную инфраструктуру для искусственного интеллекта всё изменил. Microsoft переняла стратегию Google и пошла в те же самые регионы.

Однако Чили, в которую пришла Microsoft, уже отличалась от той страны, которая когда-то радушно встречала Google. К 2022 году прошло более двух лет после Estallido Social — «Социального взрыва», навсегда изменившего политическую жизнь страны.

После месяцев протестов Чили провела необычайный демократический эксперимент. Страна попыталась переписать конституцию, сохранившуюся ещё со времён жестокой диктатуры Пиночета. В процессе активно участвовали обычные граждане. В конечном счёте новые проекты конституции не были приняты. Две разные попытки породили два идеологически односторонних документа, ни один из которых не сумел получить широкой поддержки. Но сам процесс вернул молодёжи и рабочим семьям новое чувство веры в демократию. Рост левых настроений привёл к неожиданному политическому результату: президентом был избран 35-летний левый политик нового поколения **Габриэль Борич Фонт**. В победной речи Борич повторил один из лозунгов протестов, направленный против наследия «чикагских мальчиков» времён Пиночета: **«Если Чили была колыбелью неолиберализма, то станет и его могилой».**

Борич вступил в должность в марте 2022 года. Когда-то он сам был студенческим активистом. Его победа вдохновила молодёжь по всей стране использовать опыт массовых протестов для создания нового поколения прогрессивных организаций. Эти группы постепенно становились частью повседневной жизни районов. Люди регулярно собирались и обсуждали не только конкретные проблемы, но и большие вопросы: **какой они хотят видеть будущую Чили?**

Именно тогда Арансибия вместе с другим молодым жителем Киликуры, Родриго Вальехосом, создала собственное движение. Они познакомились во время организации протестов и быстро обнаружили общую страсть к экологическим проблемам. Группу назвали: **Resistencia Socioambiental Quilicura — «Социально-экологическое сопротивление Киликуры».** Название опиралось на давно укоренившуюся в Латинской Америке идею: социальные и экологические проблемы невозможно отделить друг от друга.

Когда Microsoft пришла в Киликуру, Вальехос и Арансибия сделали то же, что до них MOSACAT и Даниэль Пенья. Начали читать всё, что компания публиковала о проекте. И проводить собственное расследование. Вальехос, студент юридического факультета Университета Диего Порталеса, ночами разбирал техническую документацию, одновременно учась тому, как вообще устроены дата-центры.

Microsoft прогнозировала значительно меньшее потребление воды, чем Google в Серрильосе. Но Вальехоса это не успокаивало. Засуха в Чили только усиливалась. По прогнозам, она могла продлиться до **2040 года**. Водно-болотные угодья Киликуры уже серьёзно страдали от наступления промышленности и ускоряющегося опустынивания. При этом на своём сайте Microsoft хвасталась новейшими технологиями охлаждения, благодаря которым дата-центры якобы вообще не должны использовать воду. Если Microsoft умеет строить дата-центры без воды, спрашивал Вальехос, **почему она не строит такой в Киликуре?** Позднее он напишет:

«Поразительно, что компания с такой репутацией, как Microsoft, публично говорит об экологической ответственности, но в реальности в стране третьего мира вроде Чили не применяет те же мировые инновационные стандарты».

Microsoft объяснит, что рекламируемые технологии ещё находятся в разработке и тестируются только на одном пилотном объекте в США. — Тогда зачем они обещают это на своём сайте? — спрашивает Вальехос. Его деятельность привлекла внимание чилийских и международных исследователей.

Среди них была Марина Отеро Верзир, директор исследовательского направления Nieuwe Instituut — Нидерландского института архитектуры, дизайна и цифровой культуры. А также исследователи Серена Дамбросио и Николас Диас Бехарано из FAIR, аналитического центра, одним из руководителей которого был Мартин Тирони Родо. Отеро поразила страсть Вальехоса и Арансибии. Но ещё сильнее — их **изнурение**. Они буквально изматывали себя, читая длинные технические документы Microsoft, публикуя критические статьи и бесконечно протестуя. Но ни компания, ни правительство почти не желали их слушать. Отеро задумалась: **как помочь им попасть за один стол с людьми, которые действительно принимают решения?** У неё было то, чего не было у Вальехоса:

связи с престижными университетами, включая Гарвард и Колумбийский университет. Такие названия заставляли Microsoft и правительство обращать внимание. Отеро начала многоуровневую кампанию и настолько погрузилась в неё, что в конце концов ушла с работы. Она выступала на крупнейших конференциях об экологическом воздействии дата-центров и борьбе Resistencia. Налаживала контакты с Министерством науки, технологий, знаний и инноваций Чили. Встречалась с представителями Microsoft и Google. Связывала Вальехоса и Арансибию с зарубежными исследователями, повышая их международную известность.

Вместе с Дамбросио и Диасом она придумала и более необычный проект. Все трое имели архитектурное образование и изучали инфраструктуру цифровых технологий через призму физической среды. И однажды они спросили себя: **а что, если смотреть на дата-центр как на архитектуру — и полностью переосмыслить его внешний вид, роль в жизни района и отношения с окружающей природой?** Диас во время путешествий любит посещать национальные библиотеки.

Это красивые здания, которые пытаются выразить величие памяти и знаний целого народа. И ему пришло в голову: ведь дата-центры тоже можно считать своего рода библиотеками. В них тоже хранятся воспоминания. В них тоже хранится знание. Почему же тогда их нельзя проектировать красивыми, открытыми и привлекательными — вместо уродливых, закрытых и экстрактивных сооружений?

Это означало полный разрыв с представлениями Microsoft и Google о том, что значит «отдавать что-то местному сообществу». Например, с программами Google, которые Диас называет

«шизофреническими»: они существуют отдельно от того реального ущерба, который инфраструктура компании наносит жителям. Вместе с Вальехосом и Арансибией исследователи нашли финансирование и организовали **четырнадцатидневный воркшоп**. Студентов-архитекторов со всего Сантьяго пригласили придумать: **каким мог бы быть дата-центр Киликуры, если бы он действительно существовал для местного сообщества?**

Получились удивительные проекты. Одна группа предложила сделать расход воды дата-центром видимым — хранить воду в больших бассейнах, которыми одновременно могли бы пользоваться жители как общественным пространством. Другая группа предложила отказаться от обычной брутальной архитектуры дата-центров и создать **«текучую территорию данных»**, где цифровая инфраструктура сосуществует с болотами и помогает уменьшать собственный вред. Конструкции дата-центра превращались бы одновременно в подвесные пешеходные дорожки. Жители Киликуры могли бы гулять среди водно-болотных угодий и наблюдать экосистему. Между обычными серверными помещениями располагались бы питомники растений и места гнездования животных, помогая восстанавливать биоразнообразие. Дата-центр забирал бы загрязнённую воду из болота, очищал её для собственных нужд, а затем возвращал обратно. Компьютеры внутри одновременно собирали бы и анализировали данные о состоянии болот, ускоряя их восстановление. — Мы не решаем проблему, — говорит Диас. — Мы просто представляем другие способы отношений между данными и водой. В конце воркшопа студенты представили проекты жителям и другим представителям общины. — Это был невероятный разговор, — вспоминает Отеро. — Сразу становилось видно, сколько знаний существует внутри самого сообщества. Людям действительно было что предложить.

На третьем году своего четырёхлетнего президентского срока Борич оказался под сильнейшим давлением. От него требовали быстрых экономических результатов — особенно учитывая, что он молод и придерживается левых взглядов. А это означало: расширять горнодобывающую промышленность;

обеспечить строительство тех самых двадцати восьми новых дата-центров. В день, когда Борич объявил о разработке национальной стратегии дата-центров, Вальехос прислал мне видео пресс-конференции и один эмодзи: **лицо со спиралями вместо глаз — выражение полного ошеломления.** Но справедливости ради, коалиция активистов с севера Чили и из Сантьяго, поддержанная исследователями внутри страны и за рубежом, всё же добилась заметных результатов.

В рамках подготовки стратегии дата-центров Министерство науки впервые создало специальный консультативный комитет из активистов. Министерство должно регулярно с ними встречаться при подготовке документа. В него пригласили Вальехоса, Арансибию и представителей MOSACAT. То же Министерство науки готовит закон об искусственном интеллекте — документ, который определит подход Чили к разработке, применению и регулированию ИИ. Раньше в министерских обсуждениях искусственный интеллект почти автоматически изображался как безусловное благо. Теперь тон изменился. Начали признавать **социальную и экологическую цену технологии**.

Чили, как и многие страны Глобального Юга, слишком многое пережила, чтобы просто ждать, пока Глобальный Север решит, каким должен быть цифровой мир. — Способ, которым мы создаём технологии, всегда отражает определённую культурную и историческую систему координат, — говорит министр Айсен Этчеверри. Глобальный интернет когда-то создавался практически без участия Чили. Теперь у страны появляется возможность участвовать в формировании искусственного интеллекта **на собственных условиях**. Мартин Тирони идёт ещё дальше.

По его словам, современная индустрия ИИ очевидным образом укоренена в **колониальной идеологии**. Она навязывает остальному миру собственное представление о том, что такое искусственный интеллект, какой ИИ является «хорошим», как должна выглядеть индустрия искусственного интеллекта. Чили могла бы стать одним из лидеров сопротивления такому навязыванию. После столетий экстрактивизма страна слишком хорошо знает, что значит видеть, как твою землю вычерпывают, отнимают и уничтожают — всё под знаменем «прогресса». Но именно этот исторический опыт может стать источником совершенно новых представлений об искусственном интеллекте. **Деколониальных представлений.**

— На мировом рынке ИИ наша страна играет совершенно определённую роль: мы даём сырьё для создания этой технологии, — говорит Тирони. — Многие компании пытаются извлечь из нас огромные объёмы ресурсов, чтобы строить искусственный интеллект. Он продолжает:

— Поэтому именно из этой точки нам и нужно мыслить — из нашего геополитического положения, из нашего положения на земле. Мы можем представить совершенно иной способ соединить технологические инновации с планетой.

Это благородная цель. Но силы, противостоящие ей, огромны.

Глава 13. Два пророка

В мае 2023 года Сэм Альтман приехал в Вашингтон, чтобы выступить перед Конгрессом. Это было почти безупречно разыгранное представление. Он снова говорил о том, что AGI — искусственный общий интеллект — сможет помочь решить проблему изменения климата и даже победить рак. Убедительно рассказывал, как технологии OpenAI не уничтожат занятость, а, напротив, создадут новые — «фантастические» — рабочие места. При этом он ловко уходил от вопросов об авторском праве, непрозрачности обучающих данных и отсутствии твёрдых гарантий конфиденциальности. И одновременно искренне призывал к регулированию. Правда, к такому регулированию, которое устраивало бы саму OpenAI.

А чтобы законодатели не слишком замедляли развитие отрасли, Альтман напоминал им о Китае. Сенаторы были очарованы. Один особенно показательный эпизод прекрасно продемонстрировал и риторическую ловкость Альтмана, и то доверие, которое он успел завоевать. Он предложил три меры государственной политики. Первая — создать специальное ведомство, которое будет лицензировать модели, превышающие определённый уровень возможностей. Вторая — разработать стандарты безопасности ИИ, позволяющие измерять его «опасные способности». Третья — обязать компании проходить независимый аудит на соответствие этим стандартам. Позже Альтман пояснил: если возможности модели трудно измерять напрямую, можно ориентироваться на объём вычислительных ресурсов, использованных при её обучении. А к «опасным способностям» он относил, например, умение модели манипулировать людьми, убеждать их или создавать инструкции для новых биологических агентов. Сенатор от Луизианы Джон Кеннеди спросил:

— Если мы примем такие правила, вы сами были бы достаточно квалифицированы, чтобы ими управлять? Альтман ответил под смех присутствующих: — Мне очень нравится моя нынешняя работа.

Кеннеди продолжил: — Но есть люди, которые подошли бы?

— Мы с удовольствием предложим вам кандидатов. Затем сенатор задал ещё один вопрос: — Вы ведь много зарабатываете? — Нет, — ответил Альтман. — Мне платят достаточно, чтобы была медицинская страховка. Доли в OpenAI у меня нет.

Рядом с ним сидел Гэри Маркус — один из самых известных критиков OpenAI. И даже Маркус в тот момент оказался впечатлён.

Он специально сказал для протокола, что никогда прежде почти не сидел так близко к Альтману и что при личном общении его искренность ощущается физически — гораздо сильнее, чем через экран телевизора. Позже Маркус от этого редкого проявления симпатии отступит. Он скажет: «Я понял, что меня, Сенат и, в конечном счёте, американский народ, вероятно, просто переиграли». Но в тот день команда, готовившая Альтмана к слушаниям, считала выступление сокрушительным успехом.

Слушания были лишь вершиной долгой кампании. После запуска ChatGPT практически все в Вашингтоне захотели встретиться с OpenAI. Небольшая политическая команда под руководством Анны Маканджу, ещё недавно работавшая почти незаметно, внезапно оказалась завалена запросами. Месяцами сотрудники OpenAI — с Альтманом и без него — завтракали, обедали и ужинали с политиками, проводили демонстрации ChatGPT, отвечали на вопросы и продвигали те самые законодательные предложения, которые Альтман затем озвучил в Конгрессе. Сенаторы. Члены Палаты представителей. Их помощники. Члены кабинета. Дипломаты. Руководители ведомств. Вице-президент Камала Харрис. Практически с утра до ночи, без остановки. К началу июня, по данным *The New York Times*, Альтман лично встретился как минимум со **ста американскими законодателями**. Некоторые из них во время слушаний даже с гордостью ссылались на свои личные разговоры с ним.

Но в тот же день в Вашингтон приехала совсем другая группа людей. Небольшая делегация голливудских концепт-художников — специалистов, которые придумывают внешний облик персонажей, предметов и целых визуальных миров для кино.

У них не было корпоративного самолёта или огромного политического отдела. Деньги на перелёт и гостиницу они собирали сообща. За публичную критику индустрии ИИ им угрожали интернет-

тролли, некоторые подверглись доксингу — их личные данные публиковали в сети. Их поездка совпала с историческим моментом: сценаристы Голливуда уже бастовали, а вскоре к ним должны были присоединиться актёры. Одной из причин забастовок была необходимость добиться защиты от использования ИИ. Художники тоже хотели рассказать законодателям, что генеративный искусственный интеллект уже делает с их профессией. Компании обучали модели на **миллионах произведений художников без их согласия**, создавали на этой основе многомиллиардные продукты — а затем эти же продукты начинали заменять самих художников. И речь шла не о горстке элитных творцов. Под угрозой находились сотни тысяч нормальных, устойчивых рабочих мест среднего класса. Художница Карла Ортис, работавшая, среди прочего, над фильмом Marvel *«Доктор Стрэндж»* и ставшая одной из истиц в первом крупном судебном процессе художников против генеративных ИИ-компаний, говорила: «Художникам сейчас невероятно тяжело. Работы нет ни у кого».

Но когда художники прибыли в Вашингтон, несколько назначенных им встреч внезапно перенесли. Причина? Выступление Сэма Альтмана. Их расписание рассыпалось. Они бродили по коридорам Конгресса — буквально **снаружи зала**, где шли слушания с Альтманом. На следующий день борьба за внимание продолжилась. Альтман отправился на закрытый ужин в Капитолии с шестьюдесятью членами Палаты представителей. Гостей угощали тщательно приготовленным буфетом с жареной курицей. Художники в то же время устроили скромный приём, стараясь заманить помощников конгрессменов всем, что могли позволить на свой бюджет: вином и едой из **Chick-fil-A**. Получилась небольшая, почти фарсовая — и одновременно довольно мрачная — иллюстрация того, **кто обладал властью и влиянием в разговоре о политике ИИ, а кто нет.**

Аргумент, которым Альтман много лет пользовался внутри OpenAI, оправдывая секретность исследований и необходимость двигаться как можно быстрее, теперь прекрасно работал и в Вашингтоне.

С его помощью дискуссию о регулировании ИИ удавалось уводить от вопросов, которые уже существовали здесь и сейчас: трудовых прав, экологических последствий, авторского права, конфиденциальности, ответственности компаний. Вместо этого внимание переносилось на будущие системы и гипотетические экстремальные угрозы. Такой подход позволял избежать жёсткой ответственности OpenAI за нынешние проблемы — а иногда даже мог укрепить её монопольное положение. Особенно эффективным оставался старый аргумент Кремниевой долины: **«А что насчёт Китая?»** Он десятилетиями помогал технологической отрасли отбиваться от регулирования. Но теперь этот козырь стал особенно сильным. Стремительный рост TikTok значительно усилил вашингтонский страх перед технологической мощью Китая.

Особенно хорошо этот настрой был заметен в Министерстве торговли США. Оно превратилось в передовой отряд американской кампании по сдерживанию китайского искусственного интеллекта. 7 октября 2022 года министерство неожиданно ввело новые экспортные ограничения. Официально целью было затормозить развитие китайских военных систем ИИ. США воспользовались механизмом **export controls — экспортного контроля**, позволяющим по соображениям национальной безопасности ограничивать продажу определённых технологий другим странам. Под ограничения попали прежде всего самые современные разработанные в США чипы для ИИ — главным образом Nvidia. Но последствия оказались куда шире, чем удар по китайскому военному сектору. Ограничения выбили почву из-под ног у всей китайской научной и технологической экосистемы ИИ: от медицинских и образовательных приложений до китайских аналогов ChatGPT. Один аналитик полупроводниковой отрасли выразился предельно резко: «Если бы пять лет назад вы рассказали мне о таких правилах, я бы сказал, что это акт войны — что мы, должно быть, уже воюем».

Однако результат разочаровал американское Министерство торговли. Китай оказался устойчивее, чем ожидалось. Темпы развития и внедрения ИИ несколько снизились — но далеко не настолько,

чтобы Вашингтон мог успокоиться. Отчасти причиной была сама Nvidia. Китай представлял для компании огромный рынок. Едва американские власти установили ограничения, Nvidia разработала новые версии чипов, которые аккуратно укладывались в разрешённые параметры, и продолжила продавать их китайским клиентам. Но санкции дали ещё один неожиданный результат. Они помогли **китайской собственной полупроводниковой индустрии**. До этого местные производители долго не могли приблизиться к качеству Nvidia. Теперь же запреты США резко повысили спрос на отечественные китайские альтернативы. В отрасль хлынули деньги, внимание и обратная связь от клиентов. То, что должно было ослабить китайскую индустрию, одновременно дало ей мощный стимул развиваться быстрее.

Но крупнейшей проблемой для американской стратегии оказалось даже не это. Им стала международная культура **открытого ИИ — open source**. По всему миру разработчики стремительно воспроизводили возможности закрытых генеративных моделей и выкладывали результаты в интернет, чтобы любой человек мог скачать их и использовать. После периода догоняющего развития Meta превратилась в одного из главных игроков этого движения. Компания активно выпускала большие языковые модели в свободный доступ. Отчасти это объяснялось убеждениями её главного научного сотрудника Янна Лекуна, давнего сторонника открытой науки. Но существовала и совершенно прагматичная коммерческая причина. Meta не обязательно было продавать сами модели, чтобы заработать. Она могла бесплатно раздавать их, одновременно встраивая ИИ в Facebook, Instagram и другие свои продукты. Так компания: укрепляла собственный статус лидера ИИ, привлекала лучших исследователей, и одновременно раздражала конкурентов, для которых продажа моделей являлась основным бизнесом.

Получив вычислительные ресурсы примерно того же масштаба, какие OpenAI имела благодаря Microsoft, Meta начала на полной скорости создавать собственный аналог серии GPT: **Llama.** Строго говоря, Llama не соответствовала классическому определению open source. Для этого Meta должна была раскрыть не только **веса модели**, но и обучающие данные. Компания сделала лишь первое: веса Llama можно было получить бесплатно. Но даже этого оказалось достаточно. Несмотря на официальную политику Meta не предоставлять Llama непосредственно в Китае, модель превратилась в один из важнейших строительных блоков китайской индустрии ИИ.

На фоне растущей тревоги Вашингтона в июле 2023 года — через два месяца после выступления Альтмана — появился программный документ, поразительно напоминавший его предложения. Его подготовила большая группа исследователей. Среди авторов были представители: OpenAI, Microsoft, Google DeepMind, а также более десятка аналитических центров, многие из которых были тесно связаны с лагерем так называемых **Doomers** — людей, считавших наиболее продвинутый ИИ потенциальной угрозой самому существованию человечества. Документ снова предлагал: ввести лицензирование ИИ-моделей по пороговым значениям вычислительной мощности; создать специальные тесты безопасности; проверять способность систем манипулировать людьми и убеждать их; оценивать возможность генерировать инструкции для создания новых биологических угроз. Но главное — пятидесятиодностраничный документ дал особое название моделям, которые, по мнению авторов, нуждались в государственном контроле: **frontier models — «передовые», или «фронтирные модели».** При этом авторы признавали: таких моделей с действительно опасными способностями **ещё не существует**. Но они могут возникнуть неожиданно — если продолжать масштабировать нынешние системы OpenAI, Anthropic и других компаний всё большим количеством вычислений.

Через несколько недель произошло ещё одно примечательное событие. OpenAI и Microsoft объединились с Google и Anthropic и создали: **Frontier Model Forum.** Организация должна была продвигать исследования безопасности наиболее передовых моделей и одновременно влиять на формирование государственной политики. Это был редкий вопрос, где сошлись интересы сразу трёх групп. **Doomers** хотели, чтобы государство сосредоточилось на угрозах будущего сверхразумного ИИ.

Boomers — технологические оптимисты — хотели продолжать быстрое развитие отрасли. А корпорациям было удобно, чтобы внимание законодателей уходило от проблем **уже существующих моделей**. Для первых речь шла об идеологии. Для вторых и третьих — ещё и о выгоде. Все они также выступали против открытой публикации весов самых передовых моделей. В первый год Meta демонстративно отсутствовала в Frontier Model Forum. Через год она всё же присоединится — чтобы тоже получить место за столом.

В основе идеи «фронтирных моделей» лежало одно ключевое предположение: **чем больше модель, тем мощнее — а значит, потенциально опаснее — её возможности.** Эта идея напрямую вытекала из философии законов масштабирования. Больше вычислений при обучении → более мощная модель. А в среде Doomers к этому добавлялось убеждение, что достаточно развитый ИИ однажды может выйти из-под контроля. Отсюда и политическая идея: регулятор должен прежде всего смотреть на то, **сколько вычислительных ресурсов было использовано для обучения модели**. Если система пересекла определённый порог — автоматически относиться к ней с большей осторожностью и вводить более жёсткие ограничения. В июльском документе 2023 года даже предложили конкретное число: **10^{26} операций с плавающей точкой.** То есть единица с двадцатью шестью нулями вычислительных операций. Если обучение модели потребовало не меньше этого объёма, её предлагалось считать «фронтирной».

Но здесь обнаружилась любопытная деталь. Сами авторы признавали: **порог во многом произволен.** По сути, его выбрали потому, что считалось, будто существующие модели ещё не пересекали этот уровень. Сара Хукер, вице-президент по исследованиям компании Cohere и одна из соавторов документа, позднее рассказывала, что после разговоров с коллегами пришла к выводу: число, вероятно, специально установили **чуть выше предполагаемого объёма вычислений, использованного OpenAI при обучении GPT-4**. Иными словами, граница между «обычной» и «опасной фронтирной» моделью оказалась проведена буквально рядом с той точкой, до которой уже дошла OpenAI.

Хукер и многие другие исследователи, включая Дебору Раджи, считали такой подход научно сомнительным. Да, большой масштаб может дать более развитые способности. Но обратное неверно: **для развитых или опасных способностей большой масштаб вовсе не обязателен.** Небольшую модель можно, например, обучить исключительно на высококачественных биологических данных — и она станет очень мощным генератором биологических инструкций. Существует и дистилляция — тот самый метод, который ещё в дорожной карте OpenAI 2021 года рассматривался как способ переносить способности больших моделей в более компактные. Большую модель можно превратить в гораздо меньшую, сохранив многие её возможности. И наоборот: само по себе увеличение масштаба не гарантирует появления конкретной способности. Всё зависит от обучающих данных. От архитектуры нейронной сети. От метода обучения. И далеко не все модели вообще построены на Transformer-архитектуре. Поэтому, утверждала Хукер, объём вычислений весьма слабо коррелирует с риском — тем более с каким-то конкретным видом риска.

Что особенно примечательно: даже среди соавторов документа не было согласия ни по поводу всей системы регулирования через вычислительные пороги, ни по поводу конкретного числа **10^{26}**. Поэтому значение спрятали в сноску и приложение, сопроводив серьёзными оговорками о несовершенстве такого подхода. Но произошло неожиданное. Именно эта цифра стремительно превратилась в одно из самых популярных предложений по регулированию ИИ в Вашингтоне. Документ идеально попал в уже существующий страх перед Китаем. Само выражение **«фронтирные модели»** звучало угрожающе. А мысль о том, что такими моделями может завладеть Пекин, — ещё страшнее. Исследовательница Эмили Вайнштейн тогда заметила: «Некоторые части администрации хватаются за всё, за что могут ухватиться, потому что хотят хоть что-нибудь сделать».

Министерство торговли охотно подхватило эту идею. Сотрудники начали консультироваться с экспертами: можно ли контролировать «фронтирные модели»? можно ли не допустить их попадания в Китай? А вскоре возникло по-настоящему беспрецедентное предложение. До сих пор экспортный контроль в основном ограничивал **оборудование** — например, передовые чипы. Теперь Министерство торговли стало размышлять: **а не запретить ли экспорт самого программного обеспечения?** То есть моделей ИИ, превышающих определённый вычислительный порог. На практике это означало бы попытку запретить свободную публикацию **весов моделей в интернете**.

При этом другое предложение, звучавшее на тех же слушаниях, почти не привлекло внимания. Гэри Маркус и вице-президент IBM Кристина Монтгомери неоднократно предлагали: **обязать компании раскрывать, на каких данных обучаются их модели.** Это не слишком помогало бы Вашингтону в противостоянии с Пекином. Зато резко усиливало бы ответственность самих корпораций. Сразу становилось бы проще разобраться: использовали ли они защищённые авторским правом произведения; какие пользовательские данные собирали; насколько научно корректны заявления о возможностях модели. Маркус говорил: «Если мы не знаем, что внутри этих моделей, мы не можем точно знать, насколько хорошо они работают». Остаётся лишь поверить компании на слово.

Раскрытие обучающих данных помогло бы и лучше понимать риски.

Почему?

Потому что поведение нейронной сети в первую очередь определяется **данными, на которых она обучалась**.

Сара Майерс Уэст, сопредседатель AI Now Institute и бывший старший советник Федеральной торговой комиссии США по ИИ, приводит простой пример.

Если большая языковая модель способна выдавать рецепты биологического оружия, то, скорее всего, причина в том, что:

её обучили на данных, содержавших информацию о биологическом оружии.

Нейронная сеть, как всегда, лишь воспроизводит найденные в обучающих данных закономерности.

Поэтому открытость данных могла бы стать первым шагом к научному пониманию того,

какие входные данные порождают опасные выходные возможности.

Пока Министерство торговли США консультировалось с экспертами и обсуждало, можно ли ограничить распространение весов моделей, информация об этих планах начала просачиваться наружу.

В начале 2024 года ведомство собиралось официально вынести вопрос на общественное обсуждение.

Но ещё до этого сама перспектива подобных ограничений расколола и сообщество разработчиков ИИ, и стремительно растущий мир специалистов по его регулированию практически пополам.

Это был спор сразу о двух вещах:

Закрытость против открытости.

И одновременно:

технологический национализм против науки без границ.

На стороне **Closed — «закрытого» подхода** оказались участники Frontier Model Forum, значительная часть сообщества Doomers, а вскоре и американские структуры национальной безопасности и разведки.

Им противостояли:

Meta,

разработчики открытого ИИ,

стартапы,

организации гражданского общества

и независимые учёные.

Закрытый лагерь продолжал использовать многие аргументы, которыми руководители OpenAI годами пользовались внутри самой компании. Сторонники открытого подхода отвечали: **изоляция моделей принесёт гораздо больше вреда, чем пользы.** Возьмём, например, аргумент о биологическом оружии. Эмили Вайнштейн отмечала: главное препятствие при создании нового биологического оружия — вовсе не отсутствие рецепта. Подобную информацию и без того можно найти в интернете, в том числе обычным поиском Google. Настоящая сложность заключается в другом: получить необходимые материалы, оборудование, лабораторную инфраструктуру и действительно изготовить опасное средство. Поэтому ограничение доступа к так называемым фронтирным моделям почти ничего не решит. А вот побочный ущерб от запрета на публикацию моделей может быть огромным.

Открытый исходный код десятилетиями был одним из фундаментов американской технологической индустрии. Стартапы могли брать уже опубликованные программы, совершенствовать их, строить на их основе новые продукты и конкурировать даже с гораздо более крупными компаниями. Если же запретить свободное распространение весов ИИ-моделей, небольшие разработчики получат ещё меньше возможностей создавать собственные сервисы. Кому это выгодно? Прежде всего гигантам, которые уже обладают: миллиардами долларов, дата-центрами, огромными командами, собственными моделями. То есть тем же компаниям, которые входили во Frontier Model Forum. Получалось, что регулирование, объявленное мерой безопасности, могло одновременно ещё сильнее закрепить их господство.

Закрытость создавала и другую проблему: ИИ становилось бы всё труднее **независимо исследовать**. Например, Саша Луччони, Ясин Жернит и Эмма Страбелл в своих работах во многом опирались на доступ к открытым генеративным моделям, чтобы измерять углеродный след и другие экологические последствия постоянного масштабирования ИИ. Закройте модели — и подобные исследования станут значительно сложнее. То есть общество получит меньше возможностей проверить заявления компаний о собственных технологиях.

Более того, международное научное сотрудничество, которое теперь всё чаще изображали как угрозу национальной безопасности, на протяжении многих лет как раз и укрепляло американское лидерство в ИИ. США и Китай — две страны, производящие наибольшее количество исследователей и научных работ по искусственному интеллекту. И более десяти лет они одновременно являлись друг для друга **главными научными партнёрами в этой сфере**. Китайские и американские исследователи постоянно развивали идеи друг друга. Это ускоряло не только фундаментальную науку, но и множество практических применений. И выигрыш получали не только две страны — польза распространялась по всему миру. Автор приводит один особенно яркий пример: **ResNet.** Эта архитектура стала одной из самых широко используемых нейронных сетей в мире. И создали её китайские исследователи, работавшие в пекинском офисе Microsoft. ResNet впоследствии легла в основу множества систем компьютерного зрения, распознавания речи и обработки языка. Она также стала одним из ключевых компонентов первой версии **AlphaFold** от DeepMind — системы, способной предсказывать трёхмерную структуру белка по последовательности аминокислот. Это имело огромное значение для ускорения разработки лекарств и изучения болезней. Позднейшие версии AlphaFold использовали уже другую архитектуру, а дальнейшая работа над системой принесла Демису Хассабису и ещё одному ведущему исследователю DeepMind Нобелевскую премию по химии 2024 года.

И всё же осенью 2023 года идеи закрытого лагеря получили мощнейшую политическую поддержку. В октябре они появились в **президентском указе администрации Джо Байдена об искусственном интеллекте**. Документ был одним из самых длинных президентских указов в американской истории. Но, как пишет автор, он производил впечатление нескольких совершенно разных документов, буквально склеенных вместе. Потому что примерно так и произошло. Одна его

часть выросла из подготовленного Белым домом ещё в 2022 году документа: ***Blueprint for an AI Bill of Rights* — «Проект Билля о правах в эпоху ИИ».** Белый дом долго разрабатывал его вместе с организациями гражданского общества. Главный вопрос там звучал примерно так: **как развивать и использовать ИИ, не разрушая гражданские права, расовую справедливость и право на частную жизнь?** Документ призывал: привлекать к созданию ИИ широкие группы общества и независимых экспертов; бороться с дискриминацией алгоритмов, например в медицине и найме; защищать людей от сбора их данных без согласия.

Но рядом с этим появился совершенно другой блок. И он почти дословно воспроизводил идеи Альтмана и июльского документа о «фронтирных моделях». Эту часть добавили довольно поздно — после того как документ о фронтирном ИИ привлёк внимание нескольких сотрудников Белого дома, близких к идеям Doomers. В исходном документе акцент делался на будущих системах, **которые ещё даже не существуют**. В президентском указе примерно половина пространства оказалась посвящена именно им. В white paper перечислялись четыре потенциально опасные способности. Указ сохранил три: создание новых рецептов химического, биологического, радиологического и ядерного оружия; автоматизированные кибератаки; и — почти дословно — способность избегать человеческого контроля: «посредством обмана и сокрытия».

Но самым удивительным для Сары Хукер, Деборы Раджи и многих других исследователей оказалось другое. В президентский указ попала **та самая конкретная цифра**: **10^{26} операций с плавающей точкой.** Именно выше этого уровня модель нужно было сообщать американскому правительству. То есть число, которое даже авторы исходного документа признавали приблизительным, спорным и практически произвольным, вдруг стало частью реальной государственной политики.

После этого идея начала распространяться почти как вирус. К концу 2023 года её подхватила Европа. Европейские законодатели, которые уже много лет готовили **EU AI Act**, оказались ошеломлены внезапным взрывом генеративного ИИ и в спешке пытались встроить новые системы в уже почти готовый закон. Они установили собственный порог: **10^{25} операций.** То есть несколько более жёсткий. В начале 2024 года тот же подход добрался до Калифорнии. Там появился законопроект об безопасности ИИ **SB 1047**. Он снова возвращался к отметке **10^{26}**. Позднее губернатор Калифорнии Гэвин Ньюсом наложил на закон вето. Ирония заключалась в том, что OpenAI и другие разработчики моделей активно лоббировали **против** SB 1047 — из-за множества других положений, усиливавших ответственность компаний. Многие обвинили Ньюсома в том, что он уступил интересам индустрии. Но Хукер, Раджи и некоторые другие исследователи неожиданно приветствовали вето. По мнению Хукер, это давало возможность вернуть дискуссию к реальному научному консенсусу. Она сказала: «Это был шаг в правильном направлении — к тому, чтобы опираться на научный консенсус. По-прежнему остаются огромные вопросы: почему именно это число и какие именно риски вы пытаетесь предотвратить?»

Вся цепочка событий выглядела поразительно: выступление Альтмана в Сенате; публикация white paper; массированная кампания влияния; паническая реакция Вашингтона на Китай; и затем стремительное превращение вычислительного порога в норму, закреплённую в важных государственных документах США и других стран. Но эта история показывала ещё одну, более глубокую проблему: **насколько сильно ослабла независимая экспертиза в области ИИ.** В сентябре 2023 года Дебора Раджи оказалась практически единственным независимым академическим исследователем — без финансовых связей ни с индустрией, ни с сообществом Doomers, — выступавшим перед Конгрессом вместе с целой группой технологических титанов. Рядом с ней сидели: Сэм Альтман, Илон Маск, Сатья Наделла, Билл Гейтс, Марк Цукерберг, Сундар Пичаи, Джек Кларк и другие руководители технологических компаний. Это был первый **AI Insight Forum**,

организованный сенатором Чаком Шумером — начало серии чрезвычайно влиятельных встреч, которые должны были сформировать будущее американского законодательства об ИИ.

Раджи слушала, как выступающие один за другим делали грандиозные, зачастую ничем не подтверждённые заявления о невероятных преимуществах ИИ — или, наоборот, о его апокалиптических угрозах. И всё это регулярно сопровождалось удачно вставленными фразами о необходимости: **«опередить Китай».** Каждый раз при упоминании Китая сенаторы словно начинали сидеть чуть прямее. Но больше всего Раджи поразило другое: **насколько охотно политики верили всему услышанному.** И тогда до неё дошло. Люди, сидевшие рядом с ней, вместе со своими огромными политическими и лоббистскими командами уже настолько долго контролировали разговор об ИИ в Вашингтоне, что многие законодатели начали воспринимать их картину мира почти как непреложную истину. Позднее представитель Шумера открыто подтвердил прессе, что сенатор лично консультируется с Альтманом и другими руководителями OpenAI при разработке регулирования. Для Раджи это стало откровением. Она вспоминала: «Вот тогда я действительно поняла масштаб проблемы. Нам нужно гораздо больше людей, которые просто проверяют подобные заявления и говорят: “Вообще-то реальность намного сложнее”».

Вашингтон был лишь кульминацией американской части политического наступления Альтмана. Ещё в марте того же года он написал в Twitter, что собирается поездить по миру и встретиться с пользователями. Очень быстро скромная идея превратилась в гигантское путешествие по множеству городов и континентов, где Альтман фотографировался чуть ли не с каждым президентом стран G20. Тур получил собственное название: **Sam Altman’s World Tour — «Мировое турне Сэма Альтмана».**

Что любопытно, никакой великой стратегии у PR- или политической команды OpenAI изначально не существовало. Альтман просто сам выбрал несколько первых остановок и сообщил о них своим более чем полутора миллионам подписчиков в Twitter. После первых выступлений интерес начал расти лавинообразно. И тогда коммуникационной и политической командам пришлось срочно включаться. С каждым новым этапом они в спешке организовывали логистику, понимая, что отсутствие предварительного планирования почти гарантированно приведёт к новым трудностям.

В этом проявлялась старая особенность Альтмана: **он привык действовать по-своему.** Иногда его личный курс совпадал с интересами компании. Иногда — нет. Хороший пример произошёл во время запуска GPT-4. OpenAI тщательно построила всю рекламную кампанию так, чтобы подчеркнуть: GPT-4 — результат труда огромной команды. Над моделью работало более ста сотрудников. Автором официального объявления значилась просто: **OpenAI.** Но затем Альтман опубликовал твит, в котором отдельно выделил одного человека — **Якуба Пахоцкого.** Пахоцкий действительно сыграл важную роль. Но он был лишь одним из **восемнадцати руководителей проекта**. Некоторые сотрудники задумались: почему именно его вклад Альтман решил представить как особенный? Затем Альтман усилил этот жест ещё больше: именно Пахоцкого он посадил позади себя во время сенатских слушаний в Вашингтоне.

Та же неопределённость проявлялась и в управлении компанией. У OpenAI были официальные руководители и формальные процедуры. Но не всегда было ясно:

действительно ли Альтман слушает именно их? И чем на самом деле определяется стратегия компании: решениями руководства, официальными процессами, или личными отношениями и импульсивными решениями самого Альтмана?

Пока OpenAI была небольшой исследовательской организацией, такая неразбериха могла казаться терпимой. Теперь компания стремительно профессионализировалась, получала глобальное влияние и оказывалась под постоянным вниманием журналистов, политиков и судов. И хаос наверху становился гораздо опаснее. OpenAI уже состояла не только из Research и Applied. Появились крупные публичные подразделения. Юридическая команда готовила заключения и разбиралась с растущим числом судебных исков. Политическая команда расширялась по всему миру. Отдел коммуникаций должен

был говорить с публикой, правительствами, журналистами и партнёрами. Компании требовался единый голос. Но единого голоса зачастую не существовало.

В конце 2023 года *The New York Times* подала в суд на OpenAI и Microsoft, обвинив их в нарушении авторских прав из-за обучения моделей на миллионах газетных статей. В начале января юридическая команда OpenAI ответила необычно резко. Она обвинила *Times* в том, что газета: «намеренно манипулировала нашими моделями», чтобы получить доказательства для судебного процесса. И буквально в ту же неделю политическое подразделение OpenAI направило материалы в комитет по коммуникациям и цифровым технологиям Палаты лордов Великобритании. Там компания заявила, что обучать её самые передовые модели без защищённых авторским правом материалов было бы: **«невозможно».** Пресса немедленно ухватилась за слово «невозможно». OpenAI поспешила от этой формулировки отступить. Один сотрудник публичного подразделения описывал ситуацию так: «Постоянно царит какая-то невероятная путаница». Да, признавал он, часть проблем можно списать на обычные болезни роста стартапа. Но публичный масштаб и влияние OpenAI выросли гораздо быстрее, чем зрелость самой организации. Он продолжал: «Я не знаю, существует ли вообще какой-то стратегический приоритет на уровне высшего руководства. Мне кажется, люди просто принимают решения каждый сами по себе. Потом вдруг это начинает выглядеть как стратегическое решение, хотя на самом деле получилось случайно. Иногда никакого плана нет — есть просто хаос».

А пока Альтман летал по Европе, Латинской Америке, Ближнему Востоку, Азии и Африке, очаровывая тщательно подобранные аудитории студентов, инвесторов и поклонников, отсутствие ясной стратегии внутри OpenAI всё сильнее обостряло старые конфликты. Компания начала раскалываться между двумя противоположными полюсами сильнее, чем когда-либо прежде. С одной стороны находилось подразделение **Applied**. Оно стремилось как можно быстрее выпускать технологии в мир, соревнуясь с беспрецедентным количеством конкурентов. Теперь его поддерживали и быстро разраставшиеся другие отделы компании, и многие исследователи. Эти люди видели невероятный успех ChatGPT и верили: **лучший способ выполнить миссию OpenAI — создавать всё более мощные модели и как можно быстрее отдавать их людям.** На противоположной стороне находился лагерь **Safety**. Он был значительно меньше.

Его участники были разбросаны по Research, особенно концентрируясь вокруг команды политических исследований Майлза Брандейджа и команды alignment Яна Лейке. Чем меньше их становилось, тем громче они били тревогу. Они считали, что новые способности моделей очень скоро могут породить не просто обычные технические проблемы, а **экзистенциальные риски**. Некоторые исследователи, которые раньше к лагерю Safety себя не относили, теперь начинали к нему присоединяться.

Быстро растущие возможности моделей убеждали их, что ИИ действительно может когда-нибудь достичь уровня интеллекта, позволяющего обмануть людей, выйти из-под контроля и начать действовать самостоятельно. Получалась почти идеальная симметрия. Одни видели рост возможностей и говорили: **мы обязаны двигаться быстрее.** Другие смотрели на тот же самый рост и отвечали: **мы обязаны быть осторожнее, чем когда-либо.** Внутри самой OpenAI столкнулись два мировоззрения: **Boomers и Doomers.**

Раскол дошёл до самого руководства. После успехов DALL-E и ChatGPT большинство топ-менеджеров стали гораздо спокойнее относиться к выпуску новых моделей. В компании для этого появился новый термин: **iterative deployment — «итеративное развёртывание».** Если GPT-2 выпускали поэтапно, а GPT-3 сначала предоставляли через контролируемый API, то новая философия была гораздо смелее: **отдавать модели пользователям рано и часто.** И, как всегда, OpenAI нашла аргумент, почему именно этот подход является самым безопасным. Альтман и другие руководители говорили: если выпускать системы постепенно, общество и институты успеют к ним адаптироваться; сама OpenAI получит возможность испытывать технологии на реальных людях; собирать настоящую обратную связь; и улучшать продукты. Во время слушаний в Сенате Альтман говорил: «Мне кажется,

ничего хорошего не получится, если тайно построить сверхмощную систему ИИ, а потом просто разом обрушить её на мир».

Но один из руководителей OpenAI явно не двигался в общем строю. **Илья Суцкевер.** Чем мощнее становились модели и чем сильнее становилось их влияние, тем больше Суцкевер убеждался: компания должна не ослаблять меры предосторожности, а наоборот — резко усиливать их. После GPT-4 он совершил радикальный поворот. Раньше почти всё своё время Суцкевер посвящал расширению возможностей моделей. Теперь примерно половину времени он начал отдавать проблемам безопасности ИИ. Людям рядом с ним временами казалось, что он ведёт войну сам с собой. Он одновременно был и **Boomer**, и **Doomer**: ещё сильнее восхищался перспективой AGI — и ещё сильнее её боялся. Он всё больше верил, что AGI может появиться очень скоро, а затем стремительно превзойти человека и превратиться в **сверхинтеллект**.

Речь Суцкевера становилась всё более мессианской. Даже его старые друзья порой не понимали, что с ним происходит. Некоторых сотрудников это откровенно тревожило. Однажды на встрече с группой новых исследователей Суцкевер рассказывал, как OpenAI должна подготовиться к появлению AGI. Он начал: — Когда мы все переберёмся в бункер… Один исследователь перебил: — Простите… в какой ещё бункер? Суцкевер ответил совершенно серьёзно: — Мы определённо построим бункер, прежде чем выпустим AGI. По его логике, столь мощная технология неизбежно станет объектом огромного интереса правительств всего мира. Она может резко усилить геополитическую напряжённость. Поэтому ключевых учёных, работающих над системой, необходимо будет физически защищать. А затем Суцкевер добавил: — Разумеется, идти в бункер или нет — каждый сможет решить сам.

Исследователь, вспоминавший этот разговор, продолжал глубоко уважать Суцкевера — но одновременно старался держаться от него на некотором расстоянии. Он говорил: «Есть группа людей — Илья один из них, — которые верят, что создание AGI приведёт к своего рода Вознесению. Буквально к Вознесению». На этом фоне у Суцкевера и Альтмана начала формироваться новая идея.

Они хотели создать специальную команду, полностью сосредоточенную на методах **выравнивания сверхинтеллекта — superintelligence alignment**. Предполагалось, что обычные методы вроде RLHF однажды перестанут работать. Ведь если система станет умнее человека, как человек сможет надёжно контролировать её через обратную связь? Альтман называл проект: **«Манхэттенским проектом по alignment».** Первоначально даже обсуждалось создание отдельной независимой некоммерческой организации с первоначальным капиталом в **1 миллиард долларов**. Но Суцкевер не хотел уходить из OpenAI, а отдельная организация столкнулась бы с проблемой доступа к самым передовым моделям. Поэтому проект решили оставить внутри компании.

Вскоре OpenAI официально объявила о создании новой команды. Она получила название: **Superalignment — «супервыравнивание».** Компания громко пообещала выделить на проект **20% всех вычислительных мощностей**, которыми OpenAI располагала на тот момент. Руководить командой должны были Суцкевер и Ян Лейке.

На одной из внутренних встреч Суцкевер вышел перед сотрудниками, чтобы представить новый проект. Он взял микрофон. И начал ударять по нему: **Бум. Бум. Бум.** А затем сказал: «Alignment — это пылающий огонь. Superalignment — это бушующий пожар». Вскоре компания переросла свой новый офис Mayo, наполненный растениями и украшенный фонтанами. Большую часть Research перевели обратно в старое здание Pioneer. И физическое разделение стало почти символическим: **Applied и Safety теперь всё больше существовали в разных мирах.** В июле 2023 года, вскоре после публичного объявления команды Superalignment, OpenAI арендовала целый кинозал в комплексе Metreon в центре Сан-Франциско. Сотрудников пригласили посмотреть фильм **«Оппенгеймер»** — историю физика Роберта Оппенгеймера, руководившего американским Манхэттенским проектом, в рамках которого было создано первое ядерное оружие. После просмотра Альтман написал в Twitter

примерно следующее: «Я надеялся, что фильм “Оппенгеймер” вдохновит новое поколение детей становиться физиками, но с этим он явно промахнулся. Надо снять такой фильм! Думаю, “Социальная сеть” сумела сделать это для основателей стартапов».

Сравнение OpenAI с Манхэттенским проектом сопровождало компанию почти с самого рождения. Ещё в самых первых письмах Илону Маску Альтман проводил такую параллель, и впоследствии она настолько укоренилась в корпоративной культуре, что её использовали даже при вводных инструктажах для новых сотрудников.

Альтман любил эту аналогию. Он с удовольствием сообщал журналистам, что родился в один день с Оппенгеймером.

Ему нравилось и перефразировать мысль создателя атомной бомбы: **«Технология появляется потому, что её возможно создать».** Но была одна часть истории, которую Альтман обычно не добавлял. Вторую половину жизни Оппенгеймер мучился угрызениями совести и активно боролся против дальнейшего распространения оружия, которое сам помог создать.

Разные сотрудники OpenAI вкладывали в образ Манхэттенского проекта совершенно разный смысл. Для большинства это была героическая история. Небольшая группа выдающихся людей, получив огромные ресурсы, совершает технологический прорыв, меняющий историю, — причём успевает сделать это раньше опасного противника.

Для лагеря Safety аналогия звучала куда мрачнее. Она подчёркивала колоссальную ответственность OpenAI: компания собиралась ввести в мир технологию, которая, по их мнению, потенциально могла привести к исчезновению человечества.

Но сам Альтман видел в истории Манхэттенского проекта ещё и урок **PR**. За несколько лет до этого он говорил на одном мероприятии: «Мир познакомился с ядерной энергией через образ, который невозможно забыть: грибовидное облако над Японией». Альтман размышлял, почему общество в какой-то момент стало бояться науки, и допускал, что одна из причин — именно этот образ. Люди увидели технологию такой мощи и решили: возможно, существуют вещи, которыми человечеству вообще не следует обладать. И вывод Альтмана был характерным: «Образы убеждают людей сильнее, чем факты».

Для самых радикальных сторонников безопасности ИИ — людей с очень высоким **p(doom)**, то есть субъективной вероятностью того, что ИИ уничтожит человечество, — спокойный оптимизм OpenAI по отношению к истории атомной бомбы становился всё более тревожным. Компания, кстати, устраивала и другой коллективный кинопросмотр — документального фильма **Apollo 11**, о высадке американцев на Луну. Космическая программа «Аполлон» была другой любимой аналогией Кремниевой долины. И один убеждённый Doomer недоумевал: «Если можно сказать “программа Аполлон”, зачем вообще вспоминать Манхэттенский проект? Зачем тащить за собой весь этот исторический груз?»

Особенно запомнилась ему одна сцена из *«Оппенгеймера»*. Перед испытанием Trinity — первым в истории взрывом атомной бомбы — Оппенгеймер, которого играет Киллиан Мёрфи, говорит, что вероятность уничтожить весь мир: **«близка к нулю».** Генерал Лесли Гровс, которого играет Мэтт Дэймон, переспрашивает: — Близка к нулю? Оппенгеймер отвечает: — А чего вы хотите от одной лишь теории? Гровс: — Ноль был бы неплох. Для радикального сторонника AI Safety эта сцена звучала почти как описание современного положения: «Вот ровно в такой ситуации мы сейчас и находимся. Мы пока вообще не умеем ничего надёжно рассчитывать относительно этих систем ИИ».

Тем временем сама аналогия с Манхэттенским проектом начала приобретать внутри OpenAI ещё один смысл. Прогресс больших языковых моделей постепенно замедлялся. Доступные объёмы новых данных сокращались. Вычислительные ресурсы тоже переставали давать прежние скачки. Поэтому Research всё больше переключалось на новое направление: **ИИ-агентов.** В каком-то смысле это означало возвращение к старому спору между двумя гипотезами. Первая утверждала: **для интеллекта**

достаточно языка. Вторая: интеллект должен быть **укоренён в реальном мире — grounded**, взаимодействовать с окружающей средой и получать из неё обратную связь.

Теперь многие исследователи считали, что чистый язык уже почти исчерпал возможности. Даже сочетание языка и зрения могло оказаться недостаточным. Следующий серьёзный скачок, предполагали они, потребует систем, которые способны **действовать в мире**. Не просто отвечать человеку.

А что-то делать. Ставить подцели. Выполнять последовательность операций. Наблюдать результат. Исправляться. И снова действовать.

Для OpenAI это было чрезвычайно привлекательно и с коммерческой точки зрения. ИИ-помощник, с которым можно разговаривать, — полезная вещь. Но ИИ, который способен сам: отправить письма, создать сайт,

написать и запустить код, выполнить сложную последовательность задач, — намного ценнее.

А ещё такой агент мог ускорить развитие **самой OpenAI**. Самым амбициозным проектом этого типа стал: **AI Scientist — «ИИ-учёный».** Цель была радикальной: создать агента, способного **самостоятельно заниматься научными исследованиями**. Логика выглядела так. Количество новых знаний, которые можно просто скачать из интернета или вытащить из учебников, конечно. Но настоящий учёный не только читает старые знания. Он создаёт новые — с помощью экспериментов. А что, если ИИ сможет делать то же самое? Запускать эксперимент. Смотреть, что получилось. Формулировать гипотезу. Проверять её. А затем получать знания, которых раньше не существовало ни в одной книге и ни на одном сайте.

Команда AI Scientist возникла после объединения двух прежних исследовательских направлений. Первое — команда Codex, занимавшаяся генерацией программного кода. Второе — группа, пытавшаяся решить проблему **рассуждения**. Для этого она использовала большие наборы математических задач с готовыми решениями, надеясь научить модель пошаговой логике. И кодирование, и математика казались особенно подходящими строительными блоками для автономного учёного. Программирование позволяло создавать и запускать эксперименты. Математическая логика — анализировать результаты. Руководили проектом **Якуб Пахоцкий и Шимон Сидор**. Но Альтман хотел не просто универсального автоматизированного учёного. Его особенно интересовал: **автономный исследователь искусственного интеллекта.** ИИ, который помогал бы OpenAI разрабатывать следующий ИИ. А тот — возможно — ещё более совершенный. То есть начинался потенциальный цикл, в котором технология помогала ускорять собственное развитие.

Для людей, веривших в близость AGI, проект был одновременно невероятно возбуждающим и пугающим. Если AI Scientist действительно заработает, рассуждали они, появление AGI резко ускорится. Но это означало и сокращение времени до одного из двух возможных исходов: **либо утопия, либо уничтожение человечества.** Разговоры о **«бункере»** начали возникать уже не только у Суцкевера. Некоторые исследователи всерьёз фантазировали о чём-то вроде современного Лос-Аламоса. Где-нибудь в удалённой американской пустыне создаётся защищённый комплекс. В нём живёт и работает элитная команда исследователей ИИ. Они изолированы от внешнего мира и защищены от потенциальных угроз. В 2023 году в OpenAI уже рассматривали два особенно серьёзных вида такой угрозы: **целенаправленные атаки на компанию** и — что звучало гораздо необычнее — **сам вышедший из-под контроля AGI.**

В Slack команда безопасности OpenAI опубликовала черновик новой модели угроз. Это был уже огромный шаг вперёд по сравнению с прежними временами, когда руководители компании лишь спорили, насколько серьёзно вообще нужно относиться к физической и информационной защите организации. В черновике выделялись три категории угроз: **1. Иностранные государства. 2. Конкуренты. 3. Идеологически мотивированные люди.** В третьей категории авторы дали ссылку

на статью **Элиезера Юдковского**. Юдковский был одной из самых известных фигур радикального лагеря AI Safety.

Именно он популяризировал термин **friendly AI — «дружественный ИИ»**, то есть систему, цели которой надёжно согласованы с человеческими. Кроме того, он написал чрезвычайно популярный фанфик: *Harry Potter and the Methods of Rationality — «Гарри Поттер и методы рационального мышления»*. Это гигантский текст — 122 главы и более 660 тысяч слов — в котором Гарри попадает в волшебный мир не просто как мальчик-волшебник, а как человек, обученный рациональному мышлению. Для многих читателей эта книга стала своеобразным входом сначала в движение **эффективного альтруизма**, а затем и в более широкую идеологию Doomers.

Юдковский также был одним из основателей сайта **LessWrong** — ключевого интеллектуального центра сообщества AI Safety. И с годами его предупреждения становились всё радикальнее. Собственную вероятность гибели человечества из-за ИИ — **p(doom)** — он оценивал уже примерно в: **95 процентов.** Юдковский начал требовать полного прекращения дальнейшего развития мощного ИИ. В марте 2023 года он опубликовал в журнале *Time* статью, где рассказывал о том, как переживал момент, когда у его дочери выпал первый молочный зуб. Он смотрел на неё и думал: **успеет ли она вообще вырасти?** Потому что, по его убеждению, развитие ИИ могло закончить человеческую историю раньше. И предложенные им меры соответствовали масштабу этого страха. Он призывал: остановить работу всех крупных GPU-кластеров; отслеживать продажу мощных процессоров; а если какая-то организация всё же тайно продолжит обучение гигантских моделей — в крайнем случае уничтожить такой «нелегальный дата-центр» **авиаударом**. Именно эту статью служба безопасности OpenAI включила в свою модель угроз как пример идеолога, публично допускающего применение насилия.

Это вызвало раздражение у небольшой группы сотрудников OpenAI, которые разделяли менее радикальные взгляды Юдковского. Им казалось странным: в документе Юдковский фигурирует как потенциальная угроза — а **враждебный, несогласованный AGI**, которого он всю жизнь боялся, вообще не упомянут. После внутренней обратной связи команда безопасности внесла изменения. Появилась четвёртая категория угроз: **«несогласованные системы AGI».**

Но пока сотрудники обсуждали угрозы сверхразума, у OpenAI зрел гораздо более непосредственный кризис. **Совет директоров начал разваливаться.** С начала 2023 года компанию один за другим покинули три независимых директора. И происходило это как раз в момент, когда успех ChatGPT запустил беспрецедентную гонку по созданию и коммерциализации генеративного ИИ. Первым, в феврале 2023 года, ушёл **Рид Хоффман**, прослуживший в совете пять лет. Причина — конфликт интересов. Годом раньше он вместе с бывшим сооснователем DeepMind Мустафой Сулейманом создал компанию **Inflection**. И она быстро превращалась в прямого конкурента OpenAI.

Через месяц ушла **Шивон Зилис** — доверенное лицо Илона Маска и директор Neuralink. После того как Маск разорвал отношения с OpenAI, Зилис продолжала наблюдать за компанией от его имени и в 2020 году официально вошла в совет директоров.

Некоторые руководители OpenAI давно опасались, что она может передавать Маску конфиденциальную информацию. Зилис заверяла их, что обязанности перед OpenAI и конфиденциальность стоят для неё выше личной лояльности бывшему сопредседателю компании. Но в июле 2022 года стало известно, что у неё родились близнецы от Маска — и она не сообщила об этом другим членам совета. После этого её независимость стала выглядеть весьма сомнительно. Тем не менее Альтман почему-то долго пытался удержать Зилис в совете. В октябре 2022 года он даже спрашивал у неё совета, как реагировать на раздражение Маска по поводу стремительно растущей оценки OpenAI, которая тогда уже приближалась примерно к **20 миллиардам долларов**. Маск, вспоминая, сколько денег он дал OpenAI в самом начале, написал Альтману: «Это подмена условий».

Но в марте 2023 года присутствие Зилис в совете стало уже практически невозможным. Маск зарегистрировал собственную компанию **xAI** — ещё одного прямого конкурента OpenAI.

Третьим ушёл **Уилл Хёрд** — бывший республиканский конгрессмен от Техаса и бывший сотрудник ЦРУ. Он вошёл в совет в 2021 году. Объявляя сотрудникам о его назначении, Альтман объяснял: в совете нужен человек, который сможет уравновесить либеральные взгляды, доминирующие в Кремниевой долине. После этого он даже организовал встречу Хёрда с сотрудниками OpenAI, чтобы все желающие могли задавать неудобные вопросы. И сотрудники действительно не сдерживались. Они подробно расспрашивали его о политических взглядах, включая отношение к Дональду Трампу. В июне 2023 года Хёрд покинул совет, чтобы полностью сосредоточиться на собственной президентской кампании в США. К октябрю он выйдет из гонки. После его ухода, помимо Альтмана, Брокмана и Суцкевера, в совете оставались всего **три независимых директора**: **Адам Д'Анджело, Таша Макколи** и **Хелен Тонер**. Из трёх оставшихся независимых директоров первым в совет вошёл **Адам Д'Анджело**. Когда-то он учился в одной школе-интернате Phillips Exeter с Марком Цукербергом, а затем два года занимал пост технического директора Facebook, прежде чем основать Quora. В 2014 году Quora вошла в Y Combinator — как раз в первую группу компаний после того, как Альтман стал президентом акселератора. Три года спустя, за год до того как Д'Анджело вошёл в совет OpenAI, Альтман увеличил инвестиции YC в Quora, став одним из руководителей раунда финансирования на **85 миллионов долларов**. В объявлении об инвестиции Альтман называл Д'Анджело одним из: «самых умных генеральных директоров Кремниевой долины». И особенно хвалил его за то, что тот мыслит на десятилетия вперёд — качество, которое, по словам Альтмана, стало редкостью среди технологических компаний.

Таша Макколи присоединилась к совету позднее, в 2018 году. Она была предпринимательницей, соосновательницей стартапа в области телеприсутствующей робототехники и руководила компанией, занимавшейся трёхмерным моделированием городов. С Альтманом её познакомил её наставник — знаменитый специалист по вычислительным технологиям **Алан Кей**. Она также знала Холдена Карнофски. В сообществе AI Safety Макколи пользовалась большим уважением. Она входила в совет некоммерческого исследовательского **Centre for the Governance of AI**, а некоторое время — и в совет британского фонда Effective Ventures Foundation, связанного с популярным в среде эффективного альтруизма проектом *80,000 Hours*. Именно Карнофски, который тогда сам входил в совет OpenAI, выдвинул её кандидатуру. В тот момент OpenAI готовилась перейти к структуре с ограниченной прибылью, и Карнофски хотел, чтобы в совете стало больше действительно независимых директоров.

Последней из трёх, в середине 2021 года, к совету присоединилась **Хелен Тонер**. Её тоже рекомендовал Карнофски — уже в качестве своей потенциальной преемницы, поскольку его трёхлетний срок подходил к концу, а одновременно создавалась Anthropic. До того как Тонер стала специалистом по Китаю и новым технологиям, она работала вместе с Карнофски в организациях GiveWell и Open Philanthropy. К проблемам безопасности ИИ она пришла через движение эффективного альтруизма. Но со временем начала отдаляться от самого движения. Ещё до краха FTX она написала один из самых заметных постов на форуме эффективного альтруизма, где отмечала, что движение становится всё более догматичным и социально замкнутым. Она даже говорила, что всё сильнее испытывает: **«разочарование в эффективном альтруизме».** Но при этом Тонер продолжала пользоваться большим уважением в этих кругах и активно работать в более широком сообществе AI Safety. Вместе с Макколи она входила в совет Centre for the Governance of AI.

К концу лета 2023 года совет директоров OpenAI уже несколько месяцев находился в тупике. Главный спор: **кого назначить новым независимым директором?** После демонстрации GPT-4, а особенно после запуска ChatGPT, Макколи примерно год занималась разработкой новой системы управления компанией. Она разговаривала с сотрудниками OpenAI и людьми за её пределами, пытаясь понять: каким должен стать обновлённый совет, какие механизмы контроля ему необходимы, и как

превратить управление OpenAI из стартапной импровизации в более профессиональную систему. Все директора, включая самого Альтмана, сначала согласились: ставки становятся настолько высокими, что новый независимый член совета должен обладать серьёзным опытом именно в **безопасности искусственного интеллекта**.

После этого совет несколько месяцев составлял список кандидатов и в конце концов провёл собеседования с пятью людьми. Среди них был **Дэн Хендрикс** — влиятельная фигура в лагере Doomers, руководитель базирующегося в Беркли Center for AI Safety и единственный советник по безопасности ИИ в компании Илона Маска xAI. Казалось бы, оставалось просто выбрать одного из пяти. Но тут процесс застопорился. Брокман начал выдвигать возражения против различных кандидатов. Суцкевер — тоже. Альтман, как часто бывало, напрямую ни с кем не спорил. Он не говорил: **«этот человек мне не подходит».** Но и не помогал продвинуть ни одну кандидатуру вперёд. Зато несколько раз предлагал новых людей. И у всех этих кандидатов было одно общее свойство: **они принадлежали к его собственной сети связей и в той или иной форме находились в сфере его влияния — финансового или личного.**

Похожее ощущение у независимых директоров возникало и при попытках создать новые механизмы контроля. Они хотели, чтобы совет получал гораздо больше информации о том, что происходит внутри OpenAI, особенно в вопросах безопасности. Но им всё чаще казалось: для Альтмана это не приоритет. Когда Макколи только вошла в совет, он назначил её своего рода связующим звеном между советом и сотрудниками. Она регулярно проводила встречи с работниками OpenAI и устраивала открытые часы, куда любой сотрудник мог прийти и поговорить с ней. Однажды Макколи даже привезла на корпоративный выезд своего мужа — актёра **Джозефа Гордона-Левитта**, который внимательно слушал технические презентации. Но во время пандемии эти встречи постепенно сошли на нет. Позднее Макколи продолжала встречаться с отдельными сотрудниками, но открытые часы так и не восстановили.

В итоге у независимых директоров не существовало систематического способа узнавать, что происходит внутри компании. Информация доходила до них через личные связи в мире технологий и AI Safety. И, конечно, через самого Альтмана. Именно это всё сильнее начинало их тревожить. Потому что рассказы Альтмана всё чаще **не совпадали с тем, что они слышали от других**. Альтман регулярно рисовал очень благополучную картину. Но из собственных источников директора получали совсем другие сведения: что к запуску ChatGPT компания была плохо подготовлена; что после запуска внутри царил серьёзный хаос; что выпуск GPT-4 сопровождался нерешёнными вопросами безопасности; что OpenAI с беспрецедентной скоростью выбрасывает на рынок всё новые продукты, прежде чем успевает разобраться с проблемами предыдущих.

Один случай показался совету особенно вопиющим. В конце 2022 года состоялось очное заседание совета — первое из тех, что планировалось проводить ежегодно. Во время этой встречи Альтман подробно рассказывал о мощных протоколах безопасности и тестирования, созданных OpenAI для оценки GPT-4. Особенно он подчёркивал роль **Deployment Safety Board — Совета по безопасности развёртывания**. Создавалось впечатление: модель не может попасть к пользователям, пока не пройдёт установленную процедуру проверки. Но после заседания один из независимых директоров случайно разговорился с сотрудником OpenAI. И тот между делом сообщил: **процедура уже была нарушена.** Microsoft провела ограниченный запуск GPT-4 среди пользователей в Индии — **без разрешения Deployment Safety Board.** Что особенно поразило директоров: Альтман только что целый день просидел с ними в одной комнате. И ни разу не сообщил об этом нарушении.

Сам факт ограниченного выпуска не означал, что публике обязательно дали опасную систему. У директоров не было оснований считать, что произошло что-то катастрофическое. Их встревожило другое. Microsoft настолько легко нарушила процедуру. А Альтман столь же легко решил об этом не

упоминать. С точки зрения членов совета, это создавало опасный прецедент. Модели OpenAI развивались невероятно быстро. И если верить сценариям AI Safety, однажды мог наступить момент, когда нарушение процедуры выпуска уже действительно будет иметь катастрофические — возможно, даже экзистенциальные — последствия. Если Альтман так небрежно относится к нарушению правил **сейчас**, что будет, когда ставки станут гораздо выше?

Были и другие тревожные эпизоды. В марте 2023 года Альтман отправил членам совета письмо — но исключил из рассылки Адама Д'Анджело. В письме он заявил: по его мнению, Д'Анджело пора покинуть совет. Причина? Quora создавала собственного чат-бота **Poe**, и Альтман утверждал, что это создаёт конфликт интересов. Для независимых директоров аргумент выглядел неожиданным и подозрительно удобным. Тонер, Макколи и Д'Анджело время от времени задавали Альтману неприятные вопросы: о безопасности OpenAI, о реальной силе некоммерческой структуры, о механизмах контроля, и о других неудобных темах. А поскольку союзников Альтмана в совете становилось всё меньше, троим независимым директорам начало казаться, что он просто ищет повод избавиться хотя бы от одного из них.

Один из независимых директоров возразил. Если Poe действительно создаёт конфликт интересов, то он всё равно гораздо меньше тех конфликтов, которые Альтман годами терпел в случаях Рида Хоффмана и Шивон Зилис. Предложение Альтмана не прошло. Д'Анджело остался в совете.

Вскоре произошёл ещё один странный эпизод. Д'Анджело был на званом ужине и услышал от кого-то, что **OpenAI Startup Fund устроен довольно необычно**. Оказалось, что инвесторы фонда получали ранний доступ к продуктам OpenAI. Это выглядело как привилегия, которую логичнее было бы предоставлять собственным инвесторам OpenAI, а не людям, вложившимся в отдельный фонд. Сначала Д'Анджело услышал об этом один раз. Потом — второй. После этого независимые директора потребовали у Альтмана документы, объясняющие юридическую структуру фонда. Когда он наконец их предоставил, выяснилось нечто ещё более странное. Фонд был не просто необычно устроен. **Юридически его владельцем был лично Сэм Альтман.** Хотя владельцем должна была быть **OpenAI**.

Для независимых директоров отдельные эпизоды постепенно складывались в единую, тревожную картину. Шаг за шагом им начинало казаться, что Альтман: ограничивает их доступ к информации, маневрирует между людьми, контролирует то, что совет узнаёт, и создаёт ситуацию, при которой совет никогда не сможет по-настоящему его контролировать. И это было особенно парадоксально. Годами Альтман сам рассказывал всему миру, что именно совет директоров является главным механизмом защиты OpenAI. Совет способен остановить его. Совет способен возразить ему. При необходимости совет может даже: **уволить его.** Теперь Альтман активно повторял этот аргумент политикам и общественности, используя независимость совета как доказательство того, что OpenAI заслуживает доверия. Но сами директора всё меньше были уверены, что располагают информацией, необходимой для выполнения этой роли.

К осени настроение среди независимых директоров достигло самой низкой точки. Месяцы переговоров о новом члене совета не дали практически никакого результата. Но времени оставалось всё меньше. Пробелы в управлении становились всё опаснее, потому что сама компания ускорялась. OpenAI готовилась приступить к обучению:

GPT-5.

Одновременно продвигался проект: **AI Scientist.** То есть системы, которая в перспективе должна была сама помогать разрабатывать новые системы ИИ. На этом фоне совет всё сильнее ощущал: что-то нужно менять — и быстро. И вдруг, совершенно неожиданно, в начале октября Хелен Тонер получила письмо. От человека, от которого она едва ли ожидала подобного обращения. **Илья Суцкевер хотел с ней поговорить.**

Глава 14

Избавление За несколько недель до того, как Илья Суцкевер связался с Хелен Тонер, Сэм Альтман оказался в центре серьёзного PR-кризиса. На протяжении многих лет история его сестры Энни — её разрыв с семьёй и последующий уход в секс-индустрию — практически не попадала в прессу. Но теперь всё изменилось.

25 сентября 2023 года журналистка Элизабет Уэйл опубликовала в журнале *New York* большой профиль Сэма Альтмана. Впервые в крупном издании там упоминалось само существование Энни. Уэйл намеренно противопоставила их жизни. С одной стороны — Энни: постоянные проблемы со здоровьем, тяжёлое финансовое положение, отсутствие уверенности даже в том, что у неё будет крыша над головой. С другой — Сэм: дома стоимостью в миллионы долларов, роскошные автомобили и статус одного из самых знаменитых предпринимателей мира. Уэйл писала, что читатели блога Альтмана, его твитов, его *Startup Playbook* и сотен статей о нём хорошо знают его братьев Джека и Макса. А Энни словно не существовала в его публичной жизни.

Она, по выражению журналистки, «никогда не должна была попасть в этот клуб». Альтман заранее знал, что материал готовится. Перед публикацией, пришедшейся на Йом-Кипур — иудейский День искупления, — *New York* обратился к OpenAI за комментарием и проводил с компанией фактчекинг. Организовывать ответы и пытаться удержать ситуацию под контролем пришлось Ханне Вонг, ставшей вице-президентом OpenAI по коммуникациям. Внезапно она оказалась посредником между журналом и семьёй Альтмана — и сама, по сути, задавалась вопросом, почему вообще семейные дела генерального директора превратились в часть её корпоративной работы.

За день до выхода статьи, когда стало окончательно ясно, что история Энни будет опубликована, она получила письмо от Сэма. Он написал ей примерно следующее: «Привет, Энни. Раз уж скоро Йом-Кипур, я хотел извиниться и попросить прощения за одну вещь». Сэм признавал, что, когда сестра просила помощи, он чувствовал себя зажатым между несколькими позициями. С одной стороны, он хотел поддержать мнение матери. С другой — соглашался с остальными членами семьи, что Энни должна научиться финансовой самостоятельности. Одновременно он считал, что ей нужна медицинская помощь и что она с трудом справляется с жизнью. Но теперь он признавал: решение было неправильным. Он должен был просто продолжать помогать ей финансово.

И искренне просил прощения. Для Альтмана эта история стала болезненным поворотом. После выхода ChatGPT его публичное восприятие было почти исключительно восторженным. Теперь же наружу вышли самые тяжёлые и личные дела семьи. А затем ситуация стала ещё хуже. В интернете всплыли старые твиты Энни, где она обвиняла Сэма в различных формах насилия. Сообщения начали широко распространяться. Для Сэма это было тяжёлым испытанием. Но для Энни публичность стала чем-то совершенно иным — освобождением накопившейся боли. Годами, на фоне ухудшения здоровья и отсутствия стабильности, ей казалось, будто она кричит в пустоту и словно вообще вычеркнута из значимой жизни окружающих.

В 2024 году автор книги связалась с Энни, чтобы услышать её версию событий. Она также обратилась к матери Энни — Конни Гибстайн, к её братьям Максу и Джеку и через пресс-службу OpenAI — к самому Сэму. Энни охотно согласилась разговаривать. Для неё возможность рассказать свою историю была принципиально важна: она считала собственный опыт необходимым для понимания морального характера брата. Она предоставила автору обширную переписку с семьёй, медицинские и психологические документы, охватывавшие период от детства до взрослой жизни, и другие материалы, отражавшие постепенный разрыв с родными и ухудшение её положения.

Мать, Конни Гибстайн, дала лишь короткий комментарий. Она подчеркнула, что семья любит Энни и беспокоится о ней, но категорически отвергла её обвинения, назвав их «ужасными, глубоко болезненными и не соответствующими действительности». При этом она отказалась от подробного разговора и не ответила на дополнительные вопросы автора. Макс и Джек не ответили вообще. OpenAI также не стала комментировать конкретные утверждения.

В январе 2025 года, после того как Энни подала иск против Сэма, он, их мать и братья выступили с гораздо более жёстким публичным заявлением. Они отрицали обвинения и утверждали, что Энни психически нестабильна, предъявляет семье необоснованные финансовые требования, а её версии событий со временем «радикально менялись». Семья заявила, что особенно тяжело наблюдать, как она якобы отказывается от традиционного лечения и нападает на родственников, которые искренне пытаются ей помочь.

Из разговоров с Энни и предоставленных ею документов у автора сложилась сложная картина. Она подчёркивает важное ограничение: ей не были известны все мотивы решений семьи. Кроме того, некоторые обвинения Энни невозможно достоверно проверить. Особенно это касается её утверждений о сексуальном насилии со стороны Сэма в детстве. Автор прямо пишет: установить истину здесь невозможно. Но для неё история Энни всё равно стала своеобразной миниатюрой всей истории OpenAI. Она помогла автору ещё яснее увидеть, насколько сама компания является отражением и продолжением личности человека, который ею руководит. Попытки Энни добиться того, чтобы её версия событий тоже осталась в публичном поле, быстро превратились из семейного вопроса в корпоративный. Материалы о сестре, как отмечает автор, раздражали Альтмана, возможно, сильнее любой другой журналистики. И глава коммуникаций OpenAI Ханна Вонг оказалась втянута в попытки ограничить распространение этой истории. Был и другой аспект. Когда обвинения Энни впервые получили серьёзную огласку, на них обратили внимание другие руководители OpenAI.

Во многих отношениях Энни и Сэм, несмотря на разницу в девять лет, удивительно похожи. Люди, близко знавшие Сэма, нередко описывают качества, которые проявлялись и у его сестры. Она прекрасно умеет слушать. Помнит мельчайшие детали о других людях. Любит дурачиться.

Очень щедра. Быстро вызывает доверие. Как и Сэм, она прекрасно училась. Из всех детей Альтманов, кроме самого Сэма, именно Энни окончила школу Burroughs. Её учитель физики Джеймс Робл и спустя более десяти лет вспоминал её как талантливую и жизнерадостную ученицу. Затем она поступила в Университет Тафтса. Основной специальностью стала биопсихология, дополнительной — танец. После выпуска в 2016 году Энни завершила необходимые подготовительные курсы, рассчитывая в будущем поступать в медицинскую школу. Она переехала в район Сан-Франциско и получила исследовательскую должность в нейробиологической лаборатории Калифорнийского университета в Сан-Франциско.

Иными словами, до того как её жизнь начала разрушаться, Энни выглядела как обычная история перспективного, амбициозного и прекрасно образованного молодого человека, перед которым открыто множество дорог. Но затем одна за другой начались проблемы. Ещё в университете, в 2014 году, ей диагностировали тендинит ахиллова сухожилия и костную шпору. Пришлось носить ортопедический ботинок. Это оказалось только первым пунктом в длинном списке заболеваний, которые впоследствии будут причинять ей хроническую боль, ограничивать подвижность и серьёзно ухудшать качество жизни. В течение шести лет Энни столкнулась с: повторяющимся тонзиллитом; хроническими болями в области таза; обострениями тендинита, распространившегося за пределы правой лодыжки; периодами, когда ей было трудно даже просто недолго стоять; увеличивающимся числом кист яичников.

В итоге ей диагностировали синдром поликистозных яичников — PCOS. А затем произошёл удар, оказавшийся ещё тяжелее. 25 мая 2018 года внезапно умер её отец.

Сердечный приступ. Из всей семьи именно с ним Энни была особенно близка. Проблемы с психическим здоровьем сопровождали её и раньше. Ещё в юности ей поставили диагнозы генерализованного тревожного расстройства и обсессивно-компульсивного расстройства. Почти десять лет она принимала Zoloft, а уже в университете постепенно прекратила приём под наблюдением психиатра. До смерти отца её состояние заметно улучшилось. Она использовала немедикаментозные способы поддержания психического здоровья — в том числе медитацию. Одновременно пыталась разобраться с физическими проблемами через здоровое питание, органические продукты и цельную пищу. Постепенно Энни заинтересовалась альтернативной медициной. Прошла подготовку преподавателя йоги. Вернулась к давней любви к искусству. Даже написала черновик книги под

названием *The Humanual* — своего рода размышление о жизни и о том, что значит быть хорошим человеком. Но смерть отца всё разрушила. Психологическое состояние резко ухудшилось. Похоже, одновременно ускорилось и ухудшение физического здоровья. К концу 2018 года посещения врачей и специалистов превратились для неё в обычную часть жизни.

Сэм тоже публично говорил, что смерть отца была худшим событием его жизни. Это произошло всего через несколько месяцев после ухода Илона Маска с должности сопредседателя OpenAI — как раз тогда, когда Сэм взял на себя всё больше власти в организации. По словам людей, близких к нему, утрата выбила его из колеи. Некоторое время он вёл себя нестабильно и тяжело переживал случившееся. Смерть отца стала переломной точкой и в отношениях между Сэмом и Энни. А также в её отношениях с остальной семьёй. Отношения с матерью и раньше были напряжёнными. По словам Энни, семья не одобряла некоторые её решения: отказ от традиционной медицинской карьеры; прекращение приёма Zoloft; увлечение естественными методами поддержания здоровья. Отец оставался единственным человеком в семье, который поддерживал её безоговорочно. После его смерти Энни всё сильнее чувствовала себя изолированной. При этом ей отчаянно хотелось одобрения и поддержки семьи. Вместо этого начались конфликты из-за денег. И отношения подошли к точке разрыва.

После смерти отца в 2018 году Энни жила в Лос-Анджелесе. Она подрабатывала помощником по написанию текстов, затем устроилась в магазин каннабиса. По её словам, по страховке жизни отца ей досталось около **100 тысяч долларов**. Энни решила использовать деньги, чтобы всерьёз попробовать построить карьеру в искусстве и выступлениях. Она ходила на курсы комедии. Выступала на открытых микрофонах. Запустила подкаст. Переделала черновик своей книги в моноспектакль и назвала его *The HumAnnie*. Но к середине 2019 года деньги начали заканчиваться. По словам Энни, значительная часть ушла на творческие проекты, дорогую аренду жилья, медицинскую страховку, которую приходилось оплачивать самостоятельно, и всё растущие расходы на здоровье: обследования;

физиотерапию; психотерапию; поездки на Lyft и Uber к врачам. Именно следующие события Энни считает началом окончательного разрушения отношений с семьёй. В мае 2019 года она узнала, что отец оставил ей свой пенсионный счёт 401(k). Она ушла из магазина и составила план на шесть месяцев. Около **40 тысяч долларов** должны были дать ей время заняться здоровьем и попытаться вывести творческие проекты на уровень, где они начнут приносить стабильный доход. Но вскоре мать сообщила: денег Энни сейчас не получит. Как пережившая супруга, Конни сохраняла контроль над пенсионными активами мужа. Она написала дочери, что с налоговой точки зрения лучше оставить средства на счёте с отсроченным налогообложением и передать их Энни через траст, доступ к которому она получит только в возрасте **59 с половиной лет**. Мать объясняла это заботой о будущем дочери. По её словам, финансовая независимость должна дать Энни долгосрочное удовлетворение, личностный рост и безопасность. Но письмо содержало и жёсткую фразу: когда оставшееся наследство отца закончится, семья больше не намерена её финансово поддерживать. Под письмом стояли подписи: **«Мама, Сэм, Макс, Джек».**

Энни тогда находилась в очень хрупком состоянии. Записи её психотерапевта того периода показывают, что решение семьи, призванное, по их мнению, помочь ей снова встать на ноги, на деле только ухудшило её состояние. В декабре 2019 года банковский счёт Энни ушёл в минус. В том же году Сэм добился первой инвестиции Microsoft в OpenAI — **1 миллиард долларов**. Компания уже на полной скорости готовилась обучать GPT-3. А Энни, испуганная и оставшаяся практически без денег, зарегистрировалась на сервисе SeekingArrangement. Во время видеозвонка она показала мужчине грудь за сумму, которой хватило, чтобы вывести банковский счёт из отрицательного баланса.

По словам Энни, до этого она никогда не просила мать или братьев содержать её. Она не считала, что имеет право автоматически рассчитывать на их богатство. Но и не думала, что, оказавшись в настоящей чрезвычайной ситуации, останется совсем без страховки. С конца 2019-го до середины 2020 года она несколько раз обращалась к семье за финансовой помощью. Пандемия только усилила

стресс и неопределённость. После двух семейных сеансов терапии с участием Сэма и матери родственники согласились оплачивать часть её расходов на протяжении некоторого времени.

Из переписки видно, что семья действительно беспокоилась. Они опасались, что постоянная финансовая помощь может закрепить, как им казалось, вредные модели поведения. Поэтому считали, что лучший способ помочь Энни восстановить психическое здоровье — продолжать подталкивать её к финансовой самостоятельности. И мать, и вся семья позднее утверждали, что действовали на основании рекомендаций специалистов.

В мае 2020 года срок финансовой помощи подходил к концу. Энни всё ещё испытывала серьёзные трудности. Она попросила продолжить оплачивать физиотерапию и психотерапию — соответственно примерно **45 и 15 долларов за сеанс**. Семья отказала. Ей сказали, что расходы за июнь она должна покрыть самостоятельно, в том числе за счёт возвращавшегося залога за жильё. Энни собрала свои немногочисленные вещи и уехала на Гавайи. Именно там она находилась, когда отец приезжал к ней в последний раз за несколько месяцев до смерти. Она нашла работу на ферме по бартерной схеме — жильё и другие базовые вещи в обмен на относительно лёгкую физическую работу: прополку, посадку растений. Примерно в это время Сэм написал ей и попросил новый адрес. У него была необычная причина. После смерти отца он попросил каждого из братьев и сестру прислать прядь волос. Волосы собирались смешать с прахом отца и превратить в искусственные бриллианты — по одному для каждого ребёнка. Сэм объяснял, что бриллианты будут состоять в основном из углерода отца, но в каждом окажется и немного от каждого из них. Теперь бриллиант Энни был готов.

Сэм хотел его отправить. Для Энни это письмо прозвучало почти как издевательство. Ей не нужен был дорогой бриллиант. Ей нужна была уверенность, что завтра у неё будут еда и жильё. Она не могла понять: Сэм действительно не осознаёт масштаб её бедствия? Или осознаёт, но ему всё равно? Пропасть между ней, Сэмом и остальной богатой семьёй казалась уже непреодолимой. Переживая огромную боль и горе, Энни перестала с ними разговаривать.

В следующие три года — до того момента, когда с Энни связалась журналистка Элизабет Уэйл, — её жизнь, по словам автора, достигла самого тяжёлого периода. У неё не было стабильного жилья. Не было уверенности, что хватит денег на еду. Не было уверенности в том, что она сможет оплачивать лечение. Чтобы сводить концы с концами, Энни занялась секс-работой — сначала виртуальной, затем и очной. В публичном заявлении семья позднее говорила, что «многими способами пыталась поддержать Энни и помочь ей обрести стабильность». В версии самой Энни одной из таких попыток стало стремление Сэма восстановить отношения весной и летом 2021 года. По её воспоминаниям, они трижды разговаривали по телефону. Сэм говорил, что очень её любит. И предложил купить ей дом. Однако стороны по-разному описывают, что именно означало это предложение. Энни утверждает, что речь шла не о том, чтобы передать дом ей в собственность, а лишь позволить ей там жить. Она понимала это как попытку не дать ей возможности продать недвижимость. И опасалась, что таким образом Сэм и их мать смогут продолжать влиять на её решения — о здоровье, лечении и карьере. Конни Гибстайн, напротив, сообщила автору, что семья предлагала Энни именно владение домом, хотя не предоставила документы, которые могли бы это подтвердить. В более позднем публичном заявлении семья сформулировала предложение иначе: они собирались купить для Энни дом через траст, чтобы она могла в нём жить, но не могла сразу его продать. В конце концов Сэм и Энни договорились о другом. Он будет в течение года напрямую оплачивать её аренду хозяину жилья. Летом 2022 года, когда Энни не восстановила общение с семьёй, выплаты прекратились. История Энни, пишет автор, делает ещё более противоречивыми те совершенно разные портреты Сэма Альтмана, которые рисуют знающие его люди. Он может быть одновременно щедрым и действующим в собственных интересах.

Мягким и угрожающим. Благодетелем для множества людей — и источником огромной личной боли для других. Публично он производит впечатление искреннего альтруиста. Но за закрытыми дверями его поступки, по мнению автора, обнаруживают гораздо более сложный расчёт. Он способен дать человеку что-то — и затем отнять. И у некоторых остаётся ощущение, будто все они являются фигурами на огромной шахматной доске, целиком видеть которую способен только один человек. А конечная партия — сохранение его собственной власти. История Энни ставит под сомнение и ещё

одну грандиозную картину, которую Альтман и другие руководители OpenAI рисовали вокруг искусственного интеллекта: будущее всеобщего изобилия. Альтман говорил, что ИИ сможет покончить с бедностью.

Брокман рассказывал, что AGI радикально улучшит здравоохранение. Суцкевер утверждал, что появится чрезвычайно эффективная и почти бесплатная психотерапия.

Но рядом существуют другие истории: работники в Кении; активисты в Чили; и собственная сестра Альтмана, непосредственно столкнувшаяся с бедностью, проблемами здоровья и отсутствием доступа к стабильной помощи. На этом фоне, замечает автор, обещания будущего начинают звучать пусто. Все достижения искусственного интеллекта никак не облегчили отчаянное положение Энни. Возможно, считает автор, технологии даже сильнее загнали её в ловушку. Секс-работой Энни заниматься не хотела. Она называла этот вариант: **«планом Z».**

К нему она пришла лишь тогда, когда хроническая боль стала настолько сильной, что даже лёгкая работа на ферме оказалась слишком тяжёлой. Сначала она пыталась зарабатывать на искусстве через интернет. Продолжала вести подкаст. Открыла магазин на Etsy. Создала страницу Patreon. Но заработка не хватало даже на оплату телефона. При этом Энни начала замечать нечто странное. В социальных сетях её публикации практически перестали получать охват. Она сохраняла скриншоты. Иногда отзывы о её подкасте в приложении Apple внезапно исчезали, что ухудшало его видимость. По меньшей мере дважды — в Instagram и на YouTube — она видела, как количество просмотров сначала увеличивалось, а затем необъяснимым образом уменьшалось. Бывший специалист Facebook по данным и два эксперта по технологическим платформам и секс-работе рассказали автору, что теоретически причиной мог стать самый первый аккаунт Энни на SeekingArrangement в декабре 2019 года. Технологические платформы иногда используют автоматизированные системы для так называемого **теневого ограничения секс-работников**. Они могут связывать между собой: устройства; адреса электронной почты; банковские счета; и другие цифровые признаки. В результате ограничения иногда распространяются даже на аккаунты, никак напрямую не связанные с секс-работой. Автор подчёркивает, что это лишь возможное объяснение, а не установленный факт. Не видя другого выхода, Энни снова вернулась на SeekingArrangement и дополнительно открыла OnlyFans.

Так её интернет-жизнь и возможности заработка оказались ещё теснее опутаны автоматическими системами модерации. Нили Мессершмидт, в прошлом руководительница в технологической индустрии, в том числе в Sony, познакомилась с Энни вскоре после её переезда в район Сан-Франциско в 2016 году. Со временем она стала для неё почти материнской фигурой, особенно после разрыва Энни с биологической семьёй. Мессершмидт сказала автору: **«Сэм принёс ИИ в мир примерно так же, как обращался со своей младшей сестрой. Он там процветает, а его сестра проваливается сквозь трещины системы».**

В конце 2020-го и начале 2021 года, уже после разрыва с семьёй, Энни, по её словам, начала переживать тяжёлые непроизвольные воспоминания о сексуальном насилии в детстве. С июля 2021-го по январь 2022 года она проходила терапию у нового специалиста по травме на острове Мауи. Записи терапевта фиксируют кризис, связанный с этими воспоминаниями, а также серьёзные сомнения Энни относительно собственной идентичности. После пятнадцати сеансов терапевт записала предварительные диагностические выводы: генерализованное тревожное расстройство;

ПТСР;

и анамнез сексуального насилия в детстве. При этом в самих записях конкретный предполагаемый насильник не назывался. Примерно в то же время Энни регулярно звонила Мессершмидт. Та сама пережила изнасилование в девятнадцать лет и говорила автору, что ещё во время первых встреч с Энни замечала в её поведении признаки, которые ассоциировала с пережитой травмой. Например, Энни чувствовала себя неуютно рядом с некоторыми мужчинами и старалась не присутствовать на встречах, когда они находились в комнате. Во время эмоциональных разговоров Энни рассказывала Мессершмидт о внезапных воспоминаниях из детства.

По воспоминаниям Мессершмидт, Энни утверждала, что в этих воспоминаниях именно Сэм неоднократно приходил к ней в постель и совершал в отношении неё сексуализированные действия; иногда, по её словам, там присутствовал и Джек.

Автор отдельно подчёркивает: спустя десятилетия чрезвычайно трудно доказать, происходило ли предполагаемое сексуальное насилие в детстве и если да, то как именно. Она приводит мнение консультировавшего её психотерапевта: у некоторых переживших насилие людей воспоминания могут долго оставаться недоступными сознанию и проявиться после определённых триггеров — например, полового созревания, начала сексуальной жизни или нового нежелательного сексуального опыта. По этой версии, углубление Энни в секс-работу могло стать источником множества таких триггеров. Но автор подчёркивает ещё раз:

цель рассказа не в том, чтобы установить, что именно произошло в детстве Энни. Цель — восстановить, **что она переживала и во что сама верила уже взрослой**, поскольку именно эти переживания в конечном итоге подтолкнули её заговорить публично.

В ноябре 2021 года, в тот период, когда Сэм оплачивал её аренду, Энни впервые публично изложила свои обвинения. В Twitter она написала, что подвергалась сексуальному, физическому, эмоциональному, словесному, финансовому и технологическому насилию со стороны своих биологических братьев и сестёр — прежде всего Сэма Альтмана и в меньшей степени Джека Альтмана. Она также заявила, что опасается наличия других пострадавших и хочет найти людей для совместного обращения к правосудию и ради предотвращения возможного вреда в будущем. Пост почти никто не заметил.

У Энни был небольшой аккаунт с малым количеством подписчиков. А известность Сэма, напротив, стремительно росла благодаря успеху GPT-3 и запуску GitHub Copilot. Вместо поддержки Энни столкнулась с троллями, которые утверждали, что она вообще не сестра Альтмана. В следующие месяцы с ней связались два журналиста.

Но Энни ещё сомневалась, насколько подробно готова рассказывать свою историю. В сентябре 2022 года она решила говорить открыто. И снова опубликовала сообщение, на этот раз прямо назвав Сэма и Джека. Она повторила обвинения в сексуальном, физическом, эмоциональном, словесном, финансовом и технологическом насилии и добавила: **«Никогда не забыто».**

Через десять месяцев — в июле 2023-го, уже после появления ChatGPT и превращения Сэма в мировую знаменитость, — Энни получила сообщение от Элизабет Уэйл из *New York*.

В течение трёх месяцев после публикации статьи *New York* доход Энни от OnlyFans вырос более чем в десять раз: примерно со **150 долларов в месяц до более чем 1500**. В один из месяцев, совпавших с кризисом совета директоров OpenAI, доход достиг примерно **5500 долларов**. После этого Энни перестала заниматься эскортом, сохранив только виртуальную секс-работу.

Семья — или, как сама Энни теперь их называла, «её родственники» — продолжала утверждать, что на протяжении многих лет оказывала ей финансовую помощь. И действительно, пишет автор, были конкретные примеры: план финансовой поддержки после семейной терапии; оплата Сэмом аренды с середины 2021-го до середины 2022 года. Но с точки зрения Энни эта помощь пришла слишком поздно — уже после того, как отчаяние заставило её заняться секс-работой. Другие предложения, включая дом, сопровождались ограничениями и условиями, которые она не хотела принимать.

В июле 2023 года Сэм отправил Энни ещё одно сообщение — на этот раз с предложением денег, по всей видимости без условий. История началась несколькими неделями раньше. Некий интернет-пользователь написал сразу им обоим и попросил Энни подробнее объяснить обвинения. Она ответила в общей переписке и подробно описала последние годы своей жизни. В тот момент Сэм находился на пике мирового турне. 5 июня, в день, когда он вместе с Суцкевером успешно выступал в Тель-Авивском университете, в его почтовом ящике лежало письмо Энни, где она спокойными, но крайне жёсткими фразами описывала пережитое.

Сэм ответил лишь 9 июля, когда тур уже закончился. Он написал, что долго не отвечал, потому что пытался понять, что именно хочет сказать. Сэм заявил, что не хочет поддерживать с Энни

постоянных отношений и понимает, что она тоже этого не хочет. При этом он готов помогать ей деньгами и надеется, что она справится с проблемами со здоровьем. Он предложил либо восстановить их прежнюю договорённость — детали которой, по его словам, уже плохо помнил, — либо выплатить ей крупную сумму единовременно. В конце он добавил, что не согласен со многими утверждениями в её письме, но считает бессмысленным их обсуждать.

К этому моменту Энни уже не хотела денег Сэма. Она хотела получить доступ к средствам, которые, как она считала, оставил ей отец. На письмо она не ответила.

Помимо пенсионного счёта 401(k), отец оставил траст, находившийся под контролем матери. В начале 2024 года, уже имея дополнительный доход, Энни смогла нанять собственного адвоката. Она узнала, что траст отца был дополнительно профинансирован в 2023 году. По мере того как история Энни становилась всё более публичной, через адвоката Конни начались переговоры о регулярных ежемесячных выплатах Энни из траста — без дополнительных условий. Впервые Энни почувствовала, что у неё появилась реальная переговорная сила. В публичном заявлении семья позднее говорила, что рассчитывает оказывать ей ежемесячную финансовую помощь **«до конца её жизни»**.

Благодаря новым выплатам Энни впервые более чем за четыре года смогла снять стабильное жильё. Вскоре она наняла ещё одного адвоката для подготовки иска против Сэма по обвинению в сексуальном насилии в детстве. Иск планировалось подать до её тридцать первого дня рождения — 8 января 2025 года. В нём утверждалось — и семья впоследствии категорически это отрицала, — что Сэм начал сексуально насиловать её, когда ей было около трёх лет, и продолжал это делать до тех пор, пока сам уже не стал взрослым, а она ещё оставалась несовершеннолетней. В иске требовалась компенсация свыше **75 тысяч долларов**.

В октябре 2024 года, после очередной серии медицинских обследований, Энни также поставили диагноз, который, по мнению её врачей, объяснял многие хронические физические проблемы: **гипермобильный синдром Элерса — Данлоса** — генетическое нарушение соединительной ткани. Автор отмечает, что такое же заболевание диагностировано у жены Грега Брокмана.

Ещё до выхода статьи *New York* Сэм начал говорить некоторым людям, что у его сестры якобы **пограничное расстройство личности**. Он подчёркивал, что это личная и деликатная семейная проблема. Один из собеседников автора охарактеризовал эффект так: **«Он обезвредил её рассказ ещё до того, как тот был опубликован».** Пограничное расстройство личности связано с серьёзными трудностями в эмоциональной регуляции и может сопровождаться чрезвычайно интенсивными отношениями, идеализацией других людей и страхом быть оставленным. Но Энни утверждает, что такого диагноза ей никогда не ставили. И, по словам автора, этого диагноза не было ни в терапевтических записях, ни в медицинских документах, которые Энни ей предоставила. Единственное упоминание появилось в августе 2021 года в записях специалиста по травме на Мауи:

терапевт отмечала, что **проверяет возможность** такого расстройства. В окончательном диагностическом заключении она указала анамнез сексуального насилия, но пограничного расстройства личности там не было. Два психотерапевта, с которыми разговаривала автор, дополнительно подчеркнули, что это расстройство во многих случаях значительно ослабевает или проходит — самостоятельно либо благодаря подходящему лечению. В числе применяемых методов они назвали медитацию и поведенческую терапию, направленную на укрепление чувства собственной ценности. Именно к таким подходам, пишет автор, Энни и обращалась. Профессор психиатрии Гарвардской медицинской школы Блейз Агирре, лечивший тысячи пациентов с пограничным расстройством, но не изучавший конкретно случай Энни, сказал: **«Исследования дают очень позитивную картину. Прогноз при этом диагнозе считается хорошим. Подавляющему большинству людей становится лучше».**

В апреле 2024 года Ханна Вонг, которой вскоре предстояло стать директором OpenAI по коммуникациям, сама заговорила с автором о психическом здоровье Энни. На протяжении шести месяцев команда Вонг обещала организовать автору посещение офиса OpenAI и интервью с

руководителями и ключевыми сотрудниками компании. Автор неоднократно пыталась услышать позицию самой OpenAI. Но через пять месяцев компания начала менять отношение к идее. За десять дней до уже забронированного автором перелёта в Сан-Франциско OpenAI сообщила: решение отменено. В офис её не пустят. В подготовке книги компания больше участвовать не будет.

Автор всё равно приехала в Сан-Франциско. Через несколько дней она сказала одному человеку, лично знакомому и с Сэмом, и с Энни, что разговаривает с Энни. На следующий день ей неожиданно написала Ханна Вонг: «Слышала, вы в городе?» Фраза показалась автору странной: именно ради заранее обсуждавшегося посещения OpenAI она и планировала эту поездку. Они встретились в кофейне Philz Coffee в районе Mission Bay неподалёку от нового офиса, который расширяла OpenAI. После непринуждённой беседы о книге Вонг сама перевела разговор на Энни. Она сказала:

«Не думаю, что выхожу за рамки дозволенного, если скажу, что у Энни есть проблемы с психическим здоровьем». При этом автор подчёркивает: до этого момента она ещё не задавала OpenAI вопросы об Энни и сама не поднимала эту тему в разговоре. Вонг продолжила: у Энни бывают хорошие дни и очень плохие. Семья пытается найти баланс между тем, чтобы защищать её, и тем, чтобы не поддерживать, по их мнению, разрушительное поведение. Затем она особенно подчеркнула: обратите внимание — семья не выступила публично с опровержением обвинений Энни. По словам Вонг, это тоже делалось ради защиты самой Энни. Она добавила, что слышала, будто в некоторых журналистских программах обсуждали, этично ли вообще было Уэйл включать историю Энни в свой профиль Альтмана. Возможно, намекала Вонг, автору книги тоже следует об этом задуматься.

Автору стало ясно: именно ради этого сообщения Вонг и предложила встречу. До этого она лично почти не отвечала на просьбы об интервью и общалась с автором преимущественно через заместителя. То, насколько важно для неё оказалось поговорить именно об Энни, показало, какое давление история сестры оказывала на Сэма — и насколько тесно личные дела Альтмана переплетались с делами самой OpenAI. Автор увидела здесь ещё одну параллель.

По её мнению, то, как Сэм и Вонг говорили о психическом состоянии Энни, напоминало отношения между «империями ИИ» и людьми, которые пытались сопротивляться их власти. Работники по разметке данных. Активисты возле дата-центров. Теперь — Энни. Все они сталкивались с гигантским разрывом сил. Чтобы хоть кто-нибудь услышал их версию, им приходилось собирать документы, письма, свидетельства и доказательства буквально по крупицам. Иногда человек начинал чувствовать, что весь мир словно вступил против него в заговор. Автор видела похожее бессилие и гнев на лицах людей в разных странах: они тратили последние ресурсы, пытаясь оспорить красивую историю технологической индустрии, а противостояли им организации, способные распоряжаться миллиардами долларов, строить колоссальную инфраструктуру и нанимать или увольнять десятки тысяч людей. А их руководители могли несколькими спокойными фразами — на конференции; в Конгрессе; перед президентом или премьер-министром; в разговоре с журналистом — снова заглушить протест и вернуть повествование под свой контроль. Но после статьи Элизабет Уэйл Энни уже не кричала в пустоту.

Её сообщения начали распространяться. И на них обратил внимание человек, имевший огромное значение для будущего OpenAI: **Илья Суцкевер.** Это произошло как раз тогда, когда Суцкевер сам всё сильнее мучился сомнениями относительно Альтмана — и того поведения, которое он воспринимал как повторяющийся образец злоупотребления властью.

ГЛАВА 15
ГАМБИТ

Через четыре дня после того, как в журнале *New York* вышел профиль Сэма Альтмана, написанный Эмили Вейл, и за четыре дня до того, как Суцкевер отправил сообщение Хелен Тонер, независимый член совета директоров встретилась ещё с одним руководителем OpenAI — техническим директором компании Мирой Мурати.

Мурати родилась в Албании и с детства научилась сохранять спокойствие посреди хаоса. Её раннее детство пришлось на бурный переход страны от тоталитарного коммунизма к либеральному капитализму. Перемены произошли так быстро, а финансовая система была настолько незрелой, что по всей стране начали стремительно распространяться финансовые пирамиды. Затем они рухнули, вызвав массовые беспорядки и насилие. По дороге в школу Мира иногда буквально обходила воронки от взрывов. Однажды учитель сказал ей: — Если ты готова продолжать, я тоже продолжу. Родители Мурати преподавали литературу, но сама она находила утешение в точности чисел.

Сначала появилась любовь к математике. Учителя быстро заметили её способности и иногда давали задания намного сложнее школьной программы, чтобы она могла двигаться быстрее. Затем пришло увлечение естественными науками: химией, биологией, физикой. А из них вырос интерес к технологиям. Мурати училась жадно. Она прочитывала всё, что попадалось под руку: сначала собственные учебники, потом перебиралась к книгам старшей сестры. И очень любила соревнование. Математические и естественно-научные олимпиады были для неё почти идеальной средой — местом, где можно было бороться с такими же сильными сверстниками за право знать больше.

В шестнадцать лет её способности принесли ей стипендию для обучения в Канаде, в частной школе Pearson College UWC в Виктории, Британская Колумбия. С этого началось стремительное восхождение. Дартмутский колледж, где она изучала машиностроение. Работа в аэрокосмической компании. А затем первый большой карьерный прорыв — Tesla, где Мурати стала старшим продакт-менеджером Model X.

Именно в Tesla она научилась создавать сложные продукты под колоссальным давлением — договариваться между командами, которые смотрят на задачу совершенно по-разному и обладают разными областями expertise. По её собственным словам, именно тогда, во время работы над автономным вождением, её всё сильнее начал привлекать искусственный интеллект. Чем глубже она погружалась в тему, тем больше убеждалась: ИИ — не просто ещё одна технология. Это фундаментальный инструмент, который в будущем понадобится почти повсюду для решения самых сложных задач.

Позднее в подкасте Кевина Скотта она скажет: «Мне стало казаться, что, возможно, это последняя вещь, над которой человечеству вообще придётся работать». Но в ИИ Мурати перешла не сразу. Через три года она ушла из Tesla в Leap Motion, где стала вице-президентом по продукту и инженерии и занялась дополненной и виртуальной реальностью. Она мечтала изменить образование. Представляла, как школьники вращают молекулы ДНК движением руки или буквально взаимодействуют с физикой летящего мяча в виртуальном пространстве. Но компания слишком рано поставила на эту технологию. VR и AR всё ещё вызывали у слишком многих людей тошноту. В 2018 году Мурати ушла в OpenAI — тогда ещё некоммерческую организацию.

В OpenAI её карьера продолжила стремительно расти. Когда лаборатория начала превращаться в коммерческую компанию, Мурати естественным образом стала вице-президентом прикладного направления и партнёрств. Она отвечала за важнейшие отношения компании — прежде всего с Microsoft — и за растущее подразделение, превращавшее научные исследования OpenAI в продукты. Мурати была даже моложе Альтмана, Суцкевера и Брокмана. И какое-то время оставалась единственной женщиной с техническим бэкграундом среди высшего руководства. Из-за этого она периодически сталкивалась с сексизмом.

Особенно со стороны исследователей, ностальгировавших по ранней OpenAI. Некоторые из них считали её «недостаточно технической»: всё-таки инженер, а не учёный. И воспринимали её возвышение как символ того, что OpenAI уходит от серьёзной фундаментальной науки в сторону бизнеса. Но в неловком, замкнутом и переполненном тестостероном мире ИИ Мурати заметно выделялась. Она хорошо чувствовала людей. Умела слушать. И почти не демонстрировала собственного эго.

Коллеги считали её человеком, который особенно хорошо решает запутанные проблемы. В бесконечной внутренней турбулентности OpenAI Мурати умела двигать компанию вперёд: слышать разные точки зрения, но при необходимости принимать неприятное решение. Один бывший коллега говорил: — Представьте ситуацию, когда нужно принять мучительно сложное решение и нет очевидно правильного ответа. А она каким-то образом помогает этот ответ найти. И поразительно часто оказывается права.

ЧЕЛОВЕК МЕЖДУ ВСЕМИ

Среди множества ролей, которые выполняла Мурати, всё важнее становилась одна: **переводчик и мост между Альтманом и остальной компанией.** После ухода команды Амодеи Альтман попросил её руководить уже не только Applied, но и Research. В мае 2022 года она официально стала техническим директором OpenAI. Всё больше команд начали подчиняться ей. И всё чаще именно через неё сотрудники взаимодействовали с Альтманом. Если Альтман хотел изменить стратегический курс — реализовывала это Мурати. Если команда считала решение Альтмана ошибочным и хотела возразить — именно Мурати отстаивала её позицию. Даже среди высшего руководства у неё был доступ к Альтману, которого не было у большинства других. Она могла прямо сказать: — Это нереалистично.

И нередко он действительно прислушивался. Она могла и честно передать коллегам: Альтман на самом деле чего-то не хочет, даже если вслух делает вид, что согласен. Когда люди уставали пытаться получить от него прямой ответ, они шли к Мурати: **объясни, что Сэм действительно думает.** Один бывший коллега говорил: — Она просто была честной. И именно она добивалась, чтобы дела действительно делались. Но чем дольше Мурати работала рядом с Альтманом, тем чаще ей приходилось **разгребать последствия его поведения**. Если две команды спорили между собой, Альтман мог отдельно встретиться с каждой — и в личной беседе согласиться с обеими. В результате каждая сторона уходила уверенной: **Сэм на нашей стороне.** Это усиливало конфликт. Порождало путаницу.

Подрывало доверие. К этому добавлялся Брокман, который постоянно врывался в чужие проекты. В глазах Мурати он был чем-то вроде второго генерального директора — только плохого. Очень категоричного. Чрезвычайно интенсивного. И способного загнать людей до полного выгорания. Во время разработки GPT-4 сочетание стилей Альтмана и Брокмана создало настолько сильное напряжение, что несколько ключевых сотрудников команды pre-training едва не ушли из компании. А именно эта команда отвечала за сбор данных и первичное обучение каждой новой модели.

И был ещё Сатья Наделла. Иногда Мурати казалось, что у OpenAI фактически **три генеральных директора**. Настолько легко Альтман уступал интересам Microsoft. Не раз Мурати тщательно выстраивала реалистичный список обязательств, которые OpenAI могла выполнить для Microsoft. А потом Альтман приходил на встречу с руководством Microsoft — и отходил от договорённого сценария. Он соглашался на запросы, совершенно не соответствовавшие реальной дорожной карте OpenAI. После чего Мурати приходилось объяснять партнёру, почему компания не может выполнить то, что только что пообещал её собственный CEO. Совету директоров Альтман представлял ситуацию совсем иначе: проблема, мол, в том, что **Мурати не умеет продуктивно работать с главным партнёром OpenAI**. После успеха ChatGPT поведение Альтмана, по мнению Мурати, становилось всё хуже. Мировая слава резко усилила внимание к нему. Календарь заполнили бесконечные поездки.

Раньше Альтман почти всегда был полон энергии. Теперь он часто выглядел полностью измотанным. Под давлением росла его тревожность. А вместе с ней усиливались и разрушительные привычки. Он продолжал делать то, что делал много лет: **говорить человеку в лицо то, что тот хочет услышать.** Но всё чаще затем за спиной говорил о нём совсем другое. Это создавало в OpenAI всё больше путаницы и вражды. Руководители команд перенимали модель поведения сверху и начинали сталкивать собственных сотрудников друг с другом.

Но теперь возникала и более серьёзная проблема. Под давлением конкурентов Альтман всё сильнее торопил компанию с релизами. Иногда пытался обходить установленные процедуры проверки. И порой, по мнению Мурати, делал это с помощью неправды. Недавно он сообщил ей, что юридическая команда OpenAI якобы разрешила выпустить GPT-4 Turbo без проверки DSB — внутреннего органа, отвечавшего за решения о развёртывании моделей. Мурати уточнила это у Джейсона Квона, руководившего юридической службой. Квон вообще не понимал, откуда у Альтмана взялась такая информация.

Летом Мурати попыталась подробно объяснить Альтману, насколько серьёзными становятся проблемы. Она надеялась: если спокойно и честно дать обратную связь, он задумается и изменит поведение. Реакция была противоположной. Альтман фактически **заморозил её**. Перестал нормально взаимодействовать. Потребовались недели, чтобы восстановить рабочие отношения. Мурати даже пришлось специально заверить его: она никому больше не рассказывала о своей критике. Она уже видела подобное с другими руководителями. Если кто-то серьёзно возражал Альтману или бросал ему вызов, он мог быстро исключить человека из важных решений. Или постепенно начать подрывать его репутацию. Происходило это достаточно тонко. Большинство сотрудников ничего не замечало. Поэтому самой Мурати потребовались годы, чтобы понять масштаб. Но со временем почти каждый крупный руководитель OpenAI успел побывать под ударом. И накопительный эффект начал разрушать верхушку компании.

Теперь очередь дошла до Суцкевера. Некоторое время назад Якуб Пахоцкий — польский исследователь, руководивший проектом AI Scientist и подчинявшийся Суцкеверу, — начал чувствовать, что ему не хватает признания и полномочий. Он обратился к своему союзнику Брокману. Брокман — к Альтману. Альтман поддержал амбиции Пахоцкого и дал ему более высокий статус в Research. Была лишь одна проблема: **Суцкеверу он ничего об этом не сказал.**

И даже после этого Альтман не стал чётко разводить зоны ответственности Суцкевера и Пахоцкого. Каждому он говорил немного разное. В результате оба начали руководить одними и теми же исследованиями — каждый в собственном направлении. И оба не могли понять: почему договориться невозможно? Несколько месяцев Research метался из стороны в сторону. Уровень стресса приблизился к критическому. Для Суцкевера всё происходящее было особенно болезненным. Это было не только унижение со стороны Альтмана. Разрушалась его многолетняя дружба с Пахоцким — дружба, выросшая из бессонных ночей, общих побед, провалов и долгих лет совместной работы над OpenAI.

Мурати снова пришлось выступать пожарным. Сначала потребовалось просто разобраться, что вообще происходит. Затем: добиться соглашения между Суцкевером и Пахоцким; остановить Брокмана, который постоянно пытался влиять на Альтмана в пользу Пахоцкого; заставить самого Альтмана придерживаться одной линии и не противоречить решениям Мурати в личных разговорах. Но она понимала: это не решает корневую проблему. Пройдёт немного времени — и появится новый кризис среди руководства. OpenAI требовалась более сильная система управления. Больше подотчётности. Настоящие механизмы контроля над CEO.

Когда Альтман на несколько недель отстранился от Мурати после её критики, больше всего его, похоже, тревожило одно: **а не рассказала ли она всё совету директоров?** На самом деле Мурати

этого не сделала. Она хотела сначала попытаться решить проблему напрямую. Да и сам совет долгие годы не вызывал у неё большого доверия. В разные периоды его структура казалась ей сомнительной. Трое сооснователей в совете, который должен быть независимым?

Слишком много. Да и многие формально независимые директора были не настолько независимы от Альтмана. У кого-то имелись с ним финансовые связи. Кто-то пользовался его сетью контактов.

Мурати прекрасно видела, как Альтман строит влияние: инвестирует в стартапы, жертвует политикам, создаёт отношения, которые потом становятся взаимными обязательствами. В разные моменты он спрашивал и её: — Что я могу для тебя сделать? Мурати неизменно уклонялась. Она не хотела запутываться в его сети. Не хотела однажды обнаружить, что **что-то ему должна**.

Но, возможно, теперь всё-таки пришло время поговорить с советом. По крайней мере наладить более регулярный канал. В конце сентября Мурати должна была приехать в Вашингтон на ежегодный фестиваль идей журнала *The Atlantic*. Там же жила Хелен Тонер. Совет как раз искал новых директоров. И следующий человек мог либо сделать совет действительно независимее — либо ещё сильнее ослабить его. Мурати считала: Альтману нужен настоящий надзор. Она написала Тонер и предложила встретиться за кофе.

КОФЕ С ХЕЛЕН ТОНЕР

Для Тонер приглашение Мурати выглядело необычно, но не подозрительно. Тонер — член совета. Мурати — один из главных руководителей компании. Логично, что она хочет поговорить.

Они встретились 29 сентября 2023 года. Разговор поначалу был совершенно обычным. Мурати рассказывала о делах компании.

OpenAI, говорила она, заработает огромные деньги — здесь никаких проблем. Сейчас особенно важны две вещи: модель Gobi;

и новая сделка, которую компания обсуждает с Microsoft. Ещё были кадровые проблемы, связанные с взаимодействием Альтмана и Брокмана. На этот раз — история между Суцкевером и Пахоцким. Но одна фраза Мурати всё же насторожила Тонер.

Альтман настолько сильно давит на компанию, требуя всё быстрее выпускать продукты, сказала Мурати, что она начинает бояться: **в какой-то момент может случиться что-нибудь плохое.**

Куда более странное событие произошло спустя несколько дней. Тонер получила письмо от Ильи Суцкевера. За два года совместной работы в совете он **ни разу не обращался к ней лично**.

Теперь он спрашивал: может ли она встретиться завтра? 4 октября они созвонились. Суцкевер нервничал настолько сильно, что с трудом говорил. Тонер решила взять инициативу на себя.

— Для меня действительно важно, чтобы у OpenAI был сильный совет, способный нормально контролировать компанию, — сказала она. — Думаю, мы все этого хотим.

Суцкевер вдруг одновременно рассмеялся и фыркнул — крайне нехарактерно для него.

— Полностью согласен, — сказал он тоном, из которого следовало, что **далеко не все согласны**. — Все согласны, — спокойно возразила Тонер. Суцкевер закатил глаза.

Тонер объяснила: независимые директора хотят усилить совет и добавить нового человека с серьёзным опытом в безопасности ИИ. Если у Суцкевера есть другие идеи, она с удовольствием их услышит. Он сразу ухватился за эту тему.

Совету, сказал он, нужно **гораздо лучше понимать, что происходит внутри компании**. Нужно внимательнее смотреть на происходящее. Тонер спросила: — А что именно, по вашему мнению, совет должен знать?

Суцкевер замолчал. Подбирал слова. Потом сказал: — Я не решаюсь прямо ответить на этот вопрос. Если бы ответил, вы бы поняли почему. И продолжил: — На самом верхнем уровне OpenAI — очень сложная среда. Гораздо сложнее, чем кажется со стороны. После паузы добавил: — Может

быть, я пересплю с этой мыслью и пойму, какими конкретными вещами могу поделиться. А пока предложил: Тонер стоит поговорить с Мурати. У неё гораздо больше контекста.

Через неделю с небольшим Мурати снова услышала от Тонер. После очередного заседания совета та сказала ей: — Кажется, происходит что-то странное. Мурати ответила: — Вы очень наблюдательны. Они договорились снова поговорить.

15 октября состоялся звонок. Тонер начала с нейтрального вопроса: — Как идут дела? Есть ли что-то, о чём совету стоит знать? Мурати осторожно подбирала слова. — Сейчас происходит много сложного, — сказала она, почти повторяя формулировку Суцкевера. Но обо всём рассказать она пока не могла. Особенно о самых сложных вещах. Потому что если произнести их вслух — назад уже не заберёшь. Тем более сейчас, когда Альтман, по её ощущениям, чувствует свою позицию CEO невероятно уязвимой.

Это удивило Тонер. — Уязвимой? Да, совет пытался создать более сильные механизмы контроля. Но никто не собирался угрожать его должности. Тем более никто даже отдалённо не обсуждал увольнение Альтмана. Мурати объяснила: Альтман — очень тревожный человек. А когда тревога становится сильной, у него появляются **глупые идеи**. Особенно если Брокман его в этом поддерживает.

Тревога запускает и определённую модель токсичного поведения. Она почти всегда одинакова.

Если кто-то сопротивляется решению Альтмана, он говорит ему всё, что, как ему кажется, тот хочет услышать. Так он получает поддержку. Но если затем терпение заканчивается и Альтман понимает, что человек всё равно будет препятствовать ему, начинается следующий этап: он постепенно **подрывает его авторитет**, пока тот не отойдёт в сторону. Это происходит тонко. Но постоянно. И совсем недавно именно так развивалась история Суцкевера и Пахоцкого. Последствия были разрушительными. Суцкевер переживал очень тяжело.

Мурати продолжила: совету особенно важно не допустить, чтобы седьмым директором стал человек Альтмана. Также нужно внимательно следить за тем, как Microsoft развёртывает технологии OpenAI, и за работой DSB. Больше она сказать пока не могла. Если Альтман узнает, что Мурати разговаривает с Тонер, он придёт в ужас. Но токсичность на самом верху компании достигла уровня, который нельзя поддерживать бесконечно. По её оценке, в следующие **шесть-девять месяцев** что-то неизбежно должно измениться. И в конце она посоветовала: — Поговорите с Ильёй. Посмотрите, чем он готов поделиться.

СОМНЕНИЯ СУЦКЕВЕРА

Когда Суцкевер впервые написал Тонер, его тревожило сразу многое. Весь последний год, наблюдая стремительный взлёт OpenAI, он всё сильнее думал о скором появлении AGI. О том, насколько радикально это изменит мир. О том, что эти изменения могут оказаться необратимыми. И о чудовищной ответственности OpenAI: добиться такого будущего, где сверхразум принесёт невероятное изобилие — а не невероятные страдания. Но постепенно его захватила другая тревога. Он начал терять веру не только в то, что OpenAI справится с AGI. Но даже в то, что компания **должна прийти к AGI под руководством Альтмана**.

После ChatGPT карьера в OpenAI превратилась в одну из самых престижных вещей Кремниевой долины. Внутри компании резко усилились: конкуренция, политика, борьба за внимание, борьба за ресурсы. Руководители команд толкались за влияние. И, по мнению Суцкевера, Альтман только ухудшал ситуацию.

Вместо того чтобы примирять амбиции людей, он каждому говорил именно то, что тот хотел услышать — и одновременно маневрировал так, чтобы получить желаемое самому. Маленьких неправд становилось столько — а временами появлялись и крупные, — что это почти превратилось в ежедневное явление. Брокман, по мнению Суцкевера, добавлял хаоса. Давно прошли дни, когда два

первых сооснователя доверяли друг другу и вместе, запершись на бесконечные часы, мечтали о будущем OpenAI. Теперь Суцкевер видел перед собой почти худшую возможную комбинацию:

хаотичная компания без ясного направления, где люди подставляют друг друга, не располагают общей информацией и больше не доверяют друг другу. А ведь без общего доверия невозможно принимать судьбоносные решения. В его глазах это разрушало обе опоры миссии OpenAI: замедляло научный прогресс; и уничтожало способность принимать разумные решения по безопасности ИИ. Теперь поведение Альтмана било уже лично по нему.

После первого разговора с Тонер 4 октября Суцкевер почти не спал. Он считал именно Тонер самым безопасным независимым директором, к которому можно обратиться. Она постоянно говорила на заседаниях о сильном управлении и безопасности.

И из всех независимых членов совета больше всего выглядела человеком, **не находящимся в кармане у Альтмана**.

В отношении Д'Анджело и Макколи он был менее уверен. Но даже с Тонер Суцкевер боялся: сколько можно рассказать? И что будет, если Альтман узнает?

В это же время в Twitter снова начали активно распространяться обвинения Энни Альтман против брата. После статьи в *New York* в компании обсуждали их почти шёпотом. Особенно широко разошлись два старых твита.

В одном, опубликованном в ноябре 2021 года, Энни обвиняла Сэма в сексуальном, физическом, эмоциональном, словесном, финансовом и технологическом насилии. Другой, от 14 марта 2023 года, был ещё прямее. Энни утверждала, что в детстве её старший брат без согласия забирался к ней в постель, и связывала это со своими обвинениями в сексуальном насилии. Суцкевер не знал, правда ли это. Но был убеждён: какими бы ни были реальные события, детство рядом с Сэмом явно оказалось для Энни тяжёлым. И это добавляло ещё один вопрос: **насколько далеко в прошлое вообще уходят проблемные модели поведения Альтмана?**

Особенно зацепило Суцкевера слово, которое использовала Энни: **abuse — насилие, злоупотребление властью.** Именно этим словом он всё чаще описывал то, что видел сам. Как и Мурати, Суцкеверу понадобились годы, чтобы распознать модель. Хотя признаки существовали почти с самого начала. Настойчивое желание Альтмана стать CEO OpenAI — без ясного и последовательного объяснения.

Мелкая ложь ранних лет — иногда настолько бессмысленная, что непонятно, зачем вообще было лгать. Предупреждения Дарио Амодеи перед уходом в 2020 году. Тогда Суцкевер не до конца понимал выражение Амодеи: **«психологическое насилие».** Теперь, когда поведение Альтмана становилось всё хуже, а последствия — всё серьёзнее, он начал понимать, что Амодеи имел в виду. 12 и 13 октября Суцкевер отправился на выездную встречу команды Superalignment. Там в рамках командного ритуала они снова сожгли символическое чучело. А сам Суцкевер продолжал мучительно думать: что делать? 16 октября он снова позвонил Тонер. На этот раз был готов рассказать больше.

Он описал историю Пахоцкого. По его словам, Альтман мог просто честно сказать: — Я хочу, чтобы Якуб играл более важную роль. И всё. Но вместо этого Альтман, во многом при содействии Брокмана, столкнул двух исследователей друг с другом. Каждый получал неполную информацию. Каждый пытался понять, почему второй ведёт себя нелогично. Суцкевер саркастически процитировал известную фразу: «Побои будут продолжаться, пока моральный дух не улучшится». Но главная проблема заключалась в другом: каждый отдельный поступок Альтмана выглядел **слишком мелким**. По отдельности почти ничего нельзя было назвать катастрофой.

Только если посмотреть на всё сразу — начинала проявляться система. Суцкевер подчёркивал: речь не о том, что «Илью сейчас обидели».

Это всего лишь очередной пример давней модели. Есть история Амодеи. Тонер стоит поговорить и с ним. Есть обвинения сестры Альтмана. — Это тоже своего рода вопрос безопасности, если вы

понимаете, что я имею в виду, — сказал Суцкевер. Его вывод был таким: Альтман, иногда вместе с Брокманом, годами одинаково обращался с разными людьми. Манипулировал. Лгал настолько привычно, что порой произносил вещи, в которые, похоже, **сам не верил**.

При этом Суцкевер оставлял возможность отступить. Он сказал: они могут всё обсудить и решить, что ничего серьёзного предпринимать не нужно. Если так — пусть забудут, что этот разговор вообще состоялся.

Но была конкретная причина, по которой он изначально написал Тонер. На конец ноября планировалась вторая ежегодная очная встреча совета. Там наконец должны были выбрать новых директоров. И Суцкевер хотел обсудить именно это. Он уже не был уверен, что расширять совет вообще стоит. Новые люди не обязательно сделают его независимее. Даже если они не окажутся сторонниками Альтмана, им потребуются месяцы или годы, чтобы научиться распознавать его методы. Контролировать его станет не легче — а труднее. Правда, Суцкевер осторожно добавил: он не до конца уверен. И хочет понять, что видит Тонер.

Тонер признала: Альтман действительно умеет быть **скользким**. Она сама видела в профессиональной жизни, как подобное поведение руководителя может каскадом создавать проблемы по всей организации. Теперь, после ухода трёх директоров, наиболее благосклонных к Альтману, она тоже считала:

следующий шаг должен дать совету реальную возможность **требовать с него ответа**. Суцкевер расслабился. Теперь он лучше понял позицию Тонер. До этого они казались противниками: она хотела добавить директоров, он — наоборот, не добавлять. Но цель у них была одна:

создать реальные ограничения власти Альтмана. Суцкевер пошёл дальше. Он понимал, что его тревога может казаться внезапной и чрезмерной.

Но, говорил он, окно возможностей быстро закрывается. И использовал язык исследований ИИ: — Совет директоров — это как **внешнее выравнивание**, а руководство — как **внутреннее выравнивание**. Тонер пытается исправить внешнюю систему контроля. А Суцкевер считает:

сломана сама внутренняя система. Тонер задумалась. — Похоже, вы считаете, что нужно рассматривать довольно серьёзные изменения. — Да, — ответил Суцкевер. Но была проблема.

Против Альтмана не существовало одного явного, вопиющего доказательства. Скорее всего, совет посмотрит на всё и решит: сделать ничего нельзя. Затем добавит новых директоров. И момент будет упущен.

Тонер предложила альтернативы. Например: установить для Альтмана конкретные цели. Через двенадцать месяцев оценить результат. Суцкевер отверг идею. Альтман выполнит любые формальные показатели. Но это никак не изменит структурную проблему его поведения. — Может быть, Брокману стоит уйти из совета? — предложила Тонер. Суцкевер признал: это действительно помогло бы. Брокман серьёзно усиливает худшие стороны динамики Альтмана.

Но корень проблемы всё равно — **сам Альтман**. — Я думал примерно в этом направлении как о плане Б, — сказал Суцкевер. Он не произнёс вслух, что именно считает **планом А**. Но к завершению разговора был уверен: Тонер начинает понимать.

ЕЩЁ ОДИН ПОЖАР

Тем временем Мурати разбиралась с очередным кризисом. Незадолго до второго разговора с Тонер Альтман внезапно начал паниковать: ему почему-то казалось, что Microsoft недовольна OpenAI. Мурати решила разобраться и организовала встречу с руководителями Microsoft, включая главу Bing Михаила Парахина. Встреча прошла лучше, чем ожидалось. Но в процессе Мурати обнаружила знакомую проблему:

Альтман снова заранее согласился на одно из требований Microsoft, не проверив, что OpenAI реально может выполнить. Он создал у партнёра ложные ожидания. А разгребать последствия опять пришлось ей. Встреча состоялась 19 октября. На следующий день, 20 октября, Мурати снова созвонилась с Тонер — уже в третий раз.

Она сказала: собирается дать Альтману очень много обратной связи. Постоянно возникает одна и та же проблема: Альтман говорит Microsoft **«да»**. Мурати потом вынуждена говорить **«нет»**. В результате отношения с партнёром расщепляются на несколько противоречащих друг другу версий реальности.

С Суцкевером и Пахоцким происходило абсолютно то же самое. Буквально в тот день наконец удалось договориться, как именно они будут сосуществовать в Research. После соглашения Мурати и Суцкевер почти умоляли Альтмана: **не меняй нашу договорённость в личном разговоре с Якубом только потому, что хочешь сказать ему приятное.** Мурати объяснила Тонер: — Когда я разговариваю с Якубом, он слышит одно. Потом идёт к Сэму — и слышит другое. На встрече Альтман пообещал придерживаться решения. Но всё оставалось хрупким. Брокман в любой момент мог снова поговорить с ним от имени Пахоцкого — и вновь перекосить ситуацию.

А Брокман, сказала Мурати, — это вообще отдельная история. Недавно он признался ей, что во время разработки GPT-4 **пытался добиться её увольнения**. Формально при этом именно Мурати была его руководителем. Раньше она даже писала оценки его работы. Но процесс каждый раз превращался в такой скандал, что в итоге она просто перестала это делать. Позднее Мурати признается другому человеку: ей самой хотелось уволить Брокмана. Но она не могла. Он сидел в совете директоров. Когда-то она думала попросить его покинуть совет. Но не решилась. Теперь хаос разросся повсюду.

Тонер спросила: — А вы что-нибудь знаете об истории Энни? Мурати ответила: нет.

Она никогда не спрашивала Сэма. У неё нет достаточного контекста о семейной истории. Но если хотя бы **десять процентов** обвинений правда — это очень плохо. Затем Мурати сказала:

— Я сама удивляюсь, что при всём происходящем всё ещё могу настолько хорошо выполнять свою работу. Она начала записывать происходящие эпизоды. Если Тонер понадобится — готова прислать больше информации. Тонер ответила:

совету нужно сосредоточиться на том, что он реально может изменить. Не на личности Сэма. Не на попытке исправить его характер. А на более сильных структурах управления, которые будут ограничивать его власть.

И это совпадало с тем, чего сама Мурати давно пыталась добиться внутри компании. Перед завершением разговора она дала Тонер один последний совет: **«Проследите, чтобы ваша информация поступала не только от Сэма».**

«СЭМ НЕ ДОЛЖЕН ДЕРЖАТЬ ПАЛЕЦ НА КНОПКЕ»

25 октября Альтман неожиданно написал Тонер. Спросил, может ли она поговорить сегодня или завтра. Тонер не понимала, чего он хочет. Двумя днями раньше, 23 октября, она снова говорила с Суцкевером. После того как Мурати сообщила ему, что сама тоже разговаривала с Тонер, Суцкевер стал значительно откровеннее. Теперь он сформулировал позицию предельно ясно: **«Я не считаю, что Сэм — тот человек, который должен держать палец на кнопке AGI».** И добавил: перед советом возникла **«огромная возможность»** что-то с этим сделать. Затем предложил конкретный вариант: **убрать Альтмана и временно назначить генеральным директором Миру Мурати.**

Позднее, когда Тонер, Макколи и Д’Анджело начали сопоставлять свои разговоры, выяснилось: Мурати тоже когда-то сказала: **«Мне некомфортно от мысли, что именно Сэм ведёт нас к AGI».** Для независимых директоров это стало крайне серьёзным сигналом. Если два ближайших заместителя Альтмана — один со стороны Applied, другой со стороны Safety — пришли к одному и тому же выводу, значит проблема гораздо глубже, чем личный конфликт.

24 октября Тонер, Д’Анджело и Макколи отдельно обсудили усиление контроля со стороны совета. Один недостаток был очевиден: некоммерческая OpenAI фактически не имела собственной независимой юридической поддержки. Почти все юридические вопросы проходили через юристов коммерческой структуры. Трое директоров решили: некоммерческой части нужны собственные адвокаты. Они должны присутствовать на заседаниях совета. И независимо анализировать сделки и юридические конструкции, которые заключает Альтман.

Тонер даже подумала: может быть, Альтман каким-то образом узнал об этих встречах и теперь хочет прекратить разговоры? Но когда они созвонились, оказалось: речь совсем о другом. О научной статье Тонер.

СТАТЬЯ

На своей основной работе в CSET Тонер за ту неделю опубликовала три статьи. Альтман выбрал одну — самую сухую и академическую. Она была посвящена концепции политической науки под названием **costly signals — «дорогостоящие сигналы»**: тому, как государства и частные компании пытаются убедительно демонстрировать обществу свои намерения в вопросах регулирования и развития ИИ. Тонер была лишь третьим соавтором. OpenAI упоминалась на страницах 28–30 документа объёмом 65 страниц. Ни в аннотации, ни на веб-странице, ни в материалах к запуску название компании вообще не фигурировало.

Судя по посещаемости, статью почти никто не читал. Тонер даже удивилась: **как Альтман вообще её нашёл?** Она не знала, что накануне статья появилась во внутреннем Slack OpenAI, в канале #policy-research-chatter. В тех нескольких страницах сравнивались сильные и слабые стороны стратегий релизов OpenAI и Anthropic. Причём OpenAI там не только критиковали. Её хвалили за честную оценку безопасности GPT-4 и за то, что компания на шесть месяцев задержала выпуск модели ради дополнительных проверок. Но глава policy planning OpenAI Дэвид Робинсон вставил в Slack лишь три абзаца — именно те, где критиковалась OpenAI и хвалилась Anthropic.

Некоторые строки он специально выделил жирным. В них противопоставлялась «гонка ко дну», которую, по мнению авторов, подстегнул ChatGPT, и более сдержанный запуск Claude после ChatGPT.

Робинсон написал: — Кстати о докладах CSET — увидел новый. Хелен Тонер в соавторах, и сравнения OpenAI с Anthropic там довольно острые. В канале началась короткая дискуссия. Один сотрудник написал:

— Да, выглядит удивительно предвзято. Не столько критика нас — она, по-моему, жёсткая, но справедливая, — сколько совершенно некритичное отношение к Anthropic. Другой добавил: — Согласен. Довольно партийно и, я бы сказал, довольно поверхностно.

Альтман сказал Тонер: буквально через несколько часов после публикации кто-то внешний прислал в OpenAI письмо с этой статьёй. Его беспокоило, что член совета директоров публично критикует компанию именно сейчас, когда OpenAI находится под пристальным вниманием регуляторов. В июле 2023 года Федеральная торговая комиссия США начала проверку в отношении практик компании: обращения с данными, обучения моделей, безопасности, а также галлюцинаций, которые могли наносить ущерб репутации людей. Тонер писала статью ещё в мае или июне — до резкого усиления регуляторного давления. Она признала: стоило перечитать её свежим взглядом с учётом нового политического контекста.

Разговор продолжался всего пятнадцать минут. Альтман всё время говорил спокойно. В конце они договорились: Тонер напишет остальным членам совета и объяснит ситуацию. В письме она выбрала примирительный тон. Извинилась за две ошибки: во-первых, решила, что статья никого особенно не заинтересует; во-вторых, именно поэтому не перечитала её достаточно внимательно перед публикацией. Она написала: «Мы с Сэмом оба согласны, что членам совета важно иметь возможность

критиковать компанию, если мы считаем нужным, но что делать это подобным образом не следовало». Ни один другой директор ей не ответил. Казалось, вопрос закрыт.

Но параллельно происходило другое. Во время третьего разговора с Суцкевером 23 октября Тонер посоветовала ему поговорить также с Макколи и Д'Анджело.

Суцкевер всё ещё не знал, можно ли им полностью доверять. После личной встречи с Д'Анджело ему показалось, что тот значительно хуже Тонер понимает проблемные стороны поведения Альтмана. 26 октября Суцкевер разговаривал по телефону с Макколи. И всё ещё боялся сказать слишком много. Но хотел выяснить одну конкретную вещь.

После разговора Альтмана с Тонер об её статье Альтман разослал нескольким сотрудникам письмо. Он написал, что говорил с Тонер и категорически не согласен с ней по поводу возможного вреда статьи: «Мне не показалось, что мы одинаково понимаем масштаб ущерба от всего этого. Любая критика со стороны члена совета имеет огромный вес». А Суцкеверу Альтман сказал ещё прямее: **Тонер должна уйти из совета.** И будто бы Макколи согласилась с ним. Что-то здесь не сходилось. Суцкевер осторожно спросил:

— Сэм сказал, что, когда обсуждал с вами статью Хелен, вы ответили: «Очевидно, Хелен должна уйти». И что после этого он стал гораздо лучше о вас думать. Это правда?

На другом конце линии Макколи была ошеломлена. — Конечно нет. Да, Альтман действительно позвонил ей поздно вечером 24 октября. Сказал, что в статье есть критическая по отношению к OpenAI часть. Упомянул, что Д'Анджело не считает это поводом для увольнения Тонер, но всё-таки считает, что ей не следовало писать подобное. Макколи ответила Альтману: она ещё не читала статью. Поэтому ему лучше напрямую поговорить с Тонер. Она совершенно точно **не говорила ничего**, что можно было бы истолковать как требование убрать Тонер из совета.

Совпадение выглядело почти невероятным. Суцкевер и независимые директора как раз обсуждали **конкретные случаи неправды Альтмана**. И тут буквально в реальном времени возник новый пример. Причём Альтман, вероятно, никогда бы не был пойман. Он прекрасно знал: Суцкевер и Макколи почти никогда напрямую не разговаривают. Если бы Суцкевер уже не начал отдельно общаться с членами совета, версии просто никогда бы не столкнулись. Но для Суцкевера это лишь подтвердило ощущение: Альтман настолько часто лжёт и маневрирует, что рано или поздно подобное неизбежно должно было произойти.

Закончив разговор с Макколи, Суцкевер сразу позвонил Тонер. Они пришли к одному выводу: **пора трём независимым членам совета поговорить друг с другом.**

Глава 16

Тайные игры Ко вторнику, 31 октября, уже все переговорили со всеми. Тонер поговорила с Макколи. Макколи позвонила Д'Анджело. Мурати поговорила и с Д'Анджело, и с Макколи.

В тот день трое независимых директоров совета — Хелен Тонер, Таша Макколи и Адам Д'Анджело — начали почти ежедневно встречаться по видеосвязи. Они пришли к выводу, что отзывы Суцкевера и Мурати об Альтмане, а также предложение Суцкевера заменить его требуют серьёзного обсуждения. Сам Суцкевер, уже твёрдо решивший, что Альтмана необходимо уволить, в этих обсуждениях не участвовал. И он, и остальные считали, что независимые директора должны прийти к собственному выводу без его влияния. Кроме того, у Суцкевера была финансовая доля в компании, и они не хотели, чтобы это обстоятельство воздействовало на решение. Позже директора назовут Суцкеверу ещё одну причину.

К тому времени их доверие к Альтману упало настолько низко, что один из них даже задавался вопросом: а не сам ли Альтман послал Суцкевера к совету директоров, чтобы проверить их лояльность и затем избавиться от тех, кто выступит против него? Независимые директора начали систематизировать всё, что знали.

И это был далеко не первый случай, когда руководители OpenAI подобным образом описывали Альтмана. В общей сложности трое директоров услышали сходные отзывы как минимум от семи человек, находившихся всего на один-два уровня ниже Альтмана. Среди них были Суцкевер, Мурати и Амодеи — люди, отвечавшие как за безопасность, так и за другие направления работы компании. Несколько человек прямо называли поведение Альтмана жестоким и манипулятивным. Большинство говорили о его нечестности и о том, что не могли доверять его словам. К этому добавлялся длинный список проблем, которые обнаружили уже сами директора: ослабление влияния некоммерческой структуры OpenAI; то, что Альтман не сообщил совету о своём юридическом контроле над OpenAI Startup Fund; то, что он не рассказал о нарушении Microsoft процедуры DSB; его попытка вытеснить из совета Д'Анджело;

а теперь, судя по всему, и Тонер. Обвинения Энни они решили вообще не рассматривать. Вопрос касался профессиональной пригодности Сэма Альтмана как генерального директора OpenAI.

Мурати говорила, что некоторые его привычки действительно можно было бы списать на типичное поведение руководителя технологической компании. Но даже в этом случае последствия были слишком серьёзными. Она сравнила Альтмана с Илоном Маском, с которым раньше работала в Tesla. Маск мог принять решение — и объяснить, почему его принял. С Альтманом всё было иначе. Мурати нередко оставалась в догадках: действительно ли он говорит ей всю правду? Вызваны ли его резкие перемены курса разумными причинами — или за ними скрывается какой-то расчёт, о котором ей ничего не известно? Так же как он создавал раздробленность в отношениях с Microsoft, Альтман создавал раздробленность и внутри собственной команды руководителей. Одним людям он сообщал одну часть информации, другим — другую. Полной картины не получал никто. И благодаря этому вся власть оставалась у него. В сочетании с Брокманом эта система, по словам Мурати, становилась разрушительной. По многим вопросам она не соглашалась с братом и сестрой Амодеи. Но здесь, говорила она, их наблюдения оказались верными.

И всё же OpenAI была не обычной технологической компанией. Независимые директора прекрасно это понимали. Возможно, это была самая влиятельная компания искусственного интеллекта в мире, занимавшаяся разработкой технологии, которую они считали одной из наиболее судьбоносных в истории. Их особенно тревожил вопрос, сформулированный Суцкевером: **что означает тот факт, что OpenAI пытается создать AGI, если высшее руководство компании не может доверять даже базовой — не говоря уже о критически важной — информации, поступающей от собственного генерального директора?**

Но они решили провести мысленный эксперимент. Допустим, OpenAI — самая обычная технологическая компания. Допустим, она просто занимается доставкой продуктов вроде Instacart. Достаточно ли тогда поведения Альтмана для увольнения? В некоторых случаях — да. В некоторых

— нет. Однако у директоров возник и другой вопрос: **а Альтман вообще лучший человек для дальнейшего управления компанией?** Он прославился как человек стартапов. Но OpenAI стремительно взрослела.

Обладал ли он навыками и характером, необходимыми для руководства зрелой компанией? И способен ли он компенсировать ту нестабильность, которую сам же создавал?

OpenAI находилась в исключительном положении. Компания была невероятно привлекательна. Если бы совет директоров провёл полноценный поиск, он мог бы выбирать среди выдающихся руководителей с огромным опытом управления зрелыми корпорациями. Да и сам Альтман, рассуждали директора, вовсе не был незаменим в повседневной работе. Пока он разъезжал по миру, основную тяжесть ежедневного управления несла Мурати. И у неё были прекрасные отношения с Microsoft.

Пока независимые директора обсуждали ситуацию, Суцкевер начал присылать им документы и скриншоты, которые собирал вместе с Мурати. Это были длинные досье, приходившие через два самоуничтожающихся электронных письма. Аватар отправителя изображал загадочного мужчину в шляпе. Как Суцкевер и предупреждал, никакой единственной убийственной улики там не было. Но складывалась общая картина: снова и снова Альтман говорил разным людям разные вещи; снова и снова высшее руководство испытывало из-за него крайнюю степень раздражения. Скриншоты показывали, что по крайней мере ещё двое старших руководителей — не связанных с безопасностью и не входивших в уже известную совету семёрку — также отмечали склонность Альтмана обходить или игнорировать процедуры. Причём как те, что вводились ради безопасности ИИ, так и обычные административные процессы. Среди прочего совет узнал о предполагаемой попытке Альтмана обойти проверку DSB для GPT-4 Turbo, неверно передав Мурати позицию юридического отдела. Была ещё одна проблема, на которую указывала Мурати: **Альтман умел почти ничего важного не фиксировать письменно.** Большинство серьёзных разговоров он предпочитал вести устно. А если впоследствии возникал спор о договорённостях, мог заявить, что собеседники просто неправильно запомнили его слова.

Но независимым директорам нужно было думать не только об этом. Как отреагирует Microsoft? Что сделает Брокман? Как поведут себя сотрудники? Они обсуждали, следует ли поговорить ещё с одним старшим руководителем, который, по словам Суцкевера и Мурати, разделял их опасения. Стоит ли провести более широкое расследование? Нужно ли подключить руководство Microsoft? После обсуждений на каждый из этих вопросов они ответили: **нет.** Каждый новый человек, которого посвящали бы в происходящее, и каждый дополнительный день промедления увеличивали вероятность того, что Альтман обо всём узнает. А дальше, опасались они, благодаря своему умению маневрировать он сможет сделать так, что совет вообще не сумеет довести обсуждение до конца.

9 ноября, когда независимые директора уже приближались к окончательному решению, Суцкевер снова поговорил с Макколи. — Сэм сказал: «Таша по-прежнему очень поддерживает идею, чтобы Хелен ушла из совета», — сообщил он ей. Это была ещё более откровенная ложь, чем раньше. Макколи с Альтманом с тех пор вообще не разговаривала.

Впоследствии директора будут считать, что роль статьи Тонер в этой истории пресса сильно преувеличила — отчасти потому, что, как они подозревали, сам Альтман мог направлять внимание журналистов именно туда. На второй день последовавшего пятидневного кризиса совета директоров они встретятся с ним при участии посредника и перечислят случаи, когда, по их мнению, он им лгал. Именно это, скажут они, окончательно разрушило доверие. В качестве одного из примеров они напомнят, что Альтман солгал Суцкеверу, будто Макколи хотела ухода Тонер из совета.

На мгновение Альтман потеряет самообладание. Его явно поймали с поличным. — Ну… я думал, что ты могла такое сказать. Не знаю, — пробормочет он.

Директора будут поражены его дерзостью. Через несколько дней первоначальные претензии Альтмана к статье Тонер появятся в прессе.

К субботе, 11 ноября, независимые директора приняли решение. Они поступят так, как предлагал Суцкевер: **уволят Альтмана и назначат Миру Мурати временным генеральным директором.** В тот же день они сообщили об этом Суцкеверу. После этого трое директоров продолжали ежедневно встречаться между собой и регулярно сверяться с Суцкевером, оформляя документы для передачи власти. Вечером в четверг, 16 ноября, все четверо позвонили Мурати по видеосвязи. Она ответила с телефона — находилась на конференции. Услышав новости, выглядела удивлённой, но восприняла их спокойно.

— Сейчас он очень, очень параноидален, — сказала она об Альтмане. — Вы уверены, что вам комфортно с этим решением? — спросили директора. — Полностью, — ответила Мурати. Она согласилась занять новую должность и выразила уверенность, что сможет объяснить решение остальным руководителям и Microsoft. Кроме того, она собиралась подключить Джейсона Вонга, которому, по её мнению, можно было доверять и который поможет сформулировать официальное заявление.

Через несколько часов, после последнего рывка по подготовке текста объявления, оставалось сделать самое важное. **Сообщить Альтману.**

В эти напряжённые последние минуты никто из членов совета ещё не представлял, насколько чудовищно они просчитались.

Уже через несколько часов после публичного объявления в пятницу, 17 ноября, ситуация для независимых директоров стала стремительно ухудшаться. Поначалу всё выглядело относительно спокойно. Мурати даже провела продуктивный разговор с Microsoft. Но вскоре она позвонила директорам уже в совершенно другом состоянии. Альтман и Брокман, сообщила она, всем рассказывают, что увольнение было **переворотом, устроенным Суцкевером**. А поскольку сам Суцкевер крайне неудачно выступил перед сотрудниками на общем собрании, важнейшие участники происходящего начали выступать против решения совета.

Вскоре директора провели видеоконференцию с откровенно враждебно настроенным руководством. И только тогда поняли, насколько всё плохо. Директор по стратегии Джейсон Квон и вице-президент по глобальным вопросам Анна Маканджу возглавили яростное сопротивление формулировке совета, согласно которой Альтман был «не всегда откровенен». Они требовали доказательств. Но директора считали, что не могут предоставить их, не раскрыв роль Мурати. При этом некоторые люди на встрече, которые, как совет знал из досье, сами выражали сомнения в лидерстве Альтмана, теперь молчали. Чем глубже ночь, тем сильнее становилась враждебность. И два самых важных союзника совета — или, по крайней мере, те, кого директора считали союзниками, — начали менять сторону.

Первым дрогнул Суцкевер. В ту же ночь, впервые всерьёз увидев возможность краха OpenAI, он начал отступать от своей прежней решимости. OpenAI была его детищем. Его жизнью. Её гибель уничтожила бы и его. Он по-прежнему не отказывался ни от одного слова, сказанного об Альтмане. Но его намерение состояло в том, чтобы **укрепить OpenAI, а не разрушить её**. Реакция сотрудников и других руководителей ошеломила его. Он был потрясён, обижен и совершенно дезориентирован. Такого результата он не ожидал. Суцкевер начал умолять остальных членов совета пересмотреть решение.

И на протяжении всего уик-энда будет делать это всё настойчивее — и всё более мучительно. Но ещё сложнее было поведение Мурати. В ходе переговоров с руководством она поддерживала связь с советом и по частным звонкам передавала информацию. Однако той твёрдости, которую продемонстрировала, соглашаясь возглавить компанию, больше не было. Она не хотела открыто использовать собственный авторитет, чтобы поддержать решение директоров. Более того, во время всё более ожесточённых встреч руководства с советом Мурати порой говорила так, словно сама столь же озадачена происходящим, как и остальные.

Для неё восстание руководителей и сотрудников означало прямую угрозу собственному положению временного генерального директора. В течение пятницы один за другим ушли: Брокман — в знак протеста; затем его союзники Якуб Пахоцкий и Шимон Сидор; а также Александр Мадры,

профессор MIT польского происхождения и близкий знакомый Пахоцкого. Внутри компании вспыхнула ярость. Она быстро перекинулась и на инвесторов. Мурати всё сильнее сомневалась, что вообще сможет удержать организацию вместе. Она начала отступать от готовности возглавить компанию. Да, решение директоров она поддерживала. Но сама в обсуждении не участвовала. Если они хотели, чтобы у неё был хоть какой-то шанс успешно принять власть, считала Мурати, теперь именно совет должен был объяснить сотрудникам и защитить собственное решение.

Уже не будучи уверенными, что могут рассчитывать на Мурати, трое независимых директоров начали срочно искать другого временного генерального директора — или новых членов совета. Особенно активно действовал Д'Анджело — единственный среди них настоящий инсайдер Кремниевой долины. В субботу и воскресенье он сделал десятки звонков, используя всю свою огромную сеть контактов. В какой-то момент директора предложили временно возглавить OpenAI Дарио Амодеи. Амодеи отказался. При этом другим людям в тот уик-энд он казался едва ли не радостно возбуждённым происходящим. В воскресенье Д'Анджело наконец нашёл человека, согласившегося на предложение совета. Это был **Эммет Шир**, сооснователь Twitch.

Поначалу он казался готовым сотрудничать и временно принять управление компанией. Но вскоре и он начал вести себя совершенно неожиданно. Позднее *The Wall Street Journal* сообщит, что Шир учился в одном наборе Y Combinator с Альтманом и был другом и наставником Брайана Чески, сооснователя Airbnb и выпускника YC. А Чески был одним из самых близких друзей Альтмана. Весь уик-энд Чески вместе с Ридом Хоффманом, членом совета Microsoft, без конца разговаривал по телефону, успокаивая инвесторов, координируя позицию Microsoft и усиливая давление. Чески быстро связался с Широм. После этого Шир встал на сторону Альтмана. К позднему вечеру воскресенья Microsoft тоже окончательно поддержала Альтмана. После заверений Чески и Хоффмана Сатья Наделла объявил, что Альтман и Брокман переходят в Microsoft, чтобы возглавить новое подразделение передовых исследований искусственного интеллекта.

Тем временем другие лаборатории ИИ кружили вокруг OpenAI словно стервятники, готовые растащить лучших специалистов из разваливающейся компании. Мурати стало очевидно: если совет не способен убедительно объяснить и защитить своё решение, от OpenAI вскоре останется лишь пустая оболочка. Она окончательно перешла на сторону остальных руководителей. Чтобы спасти компанию, Мурати теперь хотела возвращения Альтмана. Ночью сотрудники начали составлять открытое письмо против решения совета, угрожая массово уволиться и перейти в Microsoft. **Первой подпись поставила Мурати.** Многие старшие сотрудники были преданы скорее ей, чем Альтману. Именно Мурати каждый день работала вместе с ними в самой гуще событий. Именно ей они доверяли действовать не ради себя, а ради компании.

Другие считали Альтмана незаменимым по иной причине. Его тесные связи с инвесторами и уникальная способность привлекать огромные капиталы делали его лучшим — а возможно, и единственным — человеком, способным не сорвать текущий тендер по акциям OpenAI. Этот тендер обещал сотрудникам возможность превратить их доли в реальные деньги — для некоторых речь шла о миллионах долларов. Кроме того, Альтман был нужен для будущего финансирования компании. А ещё все прекрасно знали о невероятной широте его связей и способности одним звонком сделать человеку карьеру в Кремниевой долине. Когда ключевые руководители и старшие сотрудники поддержали письмо, подписи начали множиться лавинообразно. Под утро Суцкевер, уже не видевший другого пути сохранить компанию, тоже поставил своё имя. — Без Миры, — скажет позднее один исследователь, — я не думаю, что Сэм смог бы провернуть то, что провернул. Для независимых директоров колебания Мурати выглядели как самосбывающееся пророчество. Позднее *The New York Times* раскроет её роль в первоначальном отстранении Альтмана. Перед сотрудниками Мурати будет защищать себя: «У нас с Сэмом сильные и продуктивные рабочие отношения, и я никогда не стеснялась прямо сообщать ему свои замечания. Я никогда сама не обращалась к совету директоров, чтобы жаловаться на Сэма. Однако когда отдельные члены совета напрямую спрашивали меня о нём, я отвечала — и всё сказанное мной Сэм уже знал». К утру понедельника независимые директора поняли: **они проиграли.** Мурати и Суцкевер поменяли сторону. Письмо сотрудников показывало, что

дестабилизация компании стала невыносимой. Альтман вернётся. Другого способа сохранить OpenAI уже не существовало.

Оставалось одно утешение. Продержавшись достаточно долго, совет добился того, что Альтман, похоже, тоже был готов пойти на уступки. Теперь директора сосредоточились на том, чтобы сохранить хотя бы какие-то механизмы контроля над ним: оставить хотя бы одного из прежних независимых директоров ради преемственности; назначить ещё двух действительно независимых членов совета, способных противостоять Альтману; заставить самого Альтмана согласиться на расследование. Но когда соглашение уже почти было готово, в игру решил вмешаться ещё один человек. Человек, которого OpenAI когда-то отвергла как своего бывшего сопредседателя. **Илон Маск.**

Все выходные X превратился в рассадник самых невероятных теорий и конспирологических версий происходящего в OpenAI. Для самого Маска уик-энд тоже выдался насыщенным. Он руководил запуском SpaceX и завершал подготовку иска против некоммерческой организации Media Matters из-за её доклада об антисемитском контенте в X. После этого он обратился к собственной платформе, приобретённой в апреле 2022 года в результате сделки, которую многие считали крайне неудачной. Там он защищался от обвинений в антисемитизме, провозглашал абсолютную свободу слова, нападал на «воук-толпу» и традиционные СМИ. А между делом отвечал на чужие комментарии и мемы — и сам писал об OpenAI и Альтмане.

В воскресенье, 19 ноября, днём он написал: «Я очень обеспокоен. У Ильи хороший моральный компас, и он не стремится к власти. Он не пошёл бы на настолько решительный шаг, если бы не считал его абсолютно необходимым».

Позже, примерно в два часа ночи по тихоокеанскому времени, Маск опубликовал провокационный отрывок из знаменитой сцены крещения в *Крёстном отце*. Там Майкл Корлеоне окончательно превращается из морально сомневающегося сына в безжалостного нового дона, одновременно организуя убийство глав других семей. Но во вторник днём Маск решил вмешаться куда более явно. — Мне только что прислали это письмо об OpenAI, — написал он, добавив ссылку. — Похоже, приведённые там опасения заслуживают расследования. Это было не то письмо, которое подписывали нынешние сотрудники. Оно тоже было адресовано совету директоров OpenAI.

Начиналось оно так: Мы пишем вам, чтобы выразить глубокую обеспокоенность недавними событиями в OpenAI, особенно обвинениями в неправомерном поведении Сэма Альтмана. Мы — бывшие сотрудники OpenAI, покинувшие компанию в период серьёзных потрясений и конфликтов. Теперь, когда вы сами увидели, что происходит с теми, кто осмеливается выступить против Сэма Альтмана, возможно, вы понимаете, почему многие из нас молчали, опасаясь последствий. Больше молчать мы не можем.

Далее следовала серия требований. Главное — временный генеральный директор Эммет Шир должен расширить расследование поведения Альтмана и включить в него более раннюю историю OpenAI и преобразование компании из некоммерческой структуры. В письме утверждалось: «Мы считаем, что значительное число сотрудников OpenAI было вытеснено из компании ради облегчения перехода к коммерческой модели».

Также авторы выдвигали целый ряд обвинений, описывая: «тревожащую модель обмана и манипуляций со стороны Сэма Альтмана и Грега Брокмана». В самом конце находился раздел **«Дополнительное чтение для широкой публики»** с тремя ссылками. Одна вела на ветку в X Джеффри Ирвинга, специалиста по безопасности ИИ, ушедшего в 2019 году в DeepMind. Ирвинг писал, что Альтман: «лгал мне в разных случаях»

и что по отношению к другим людям он бывал: «нечестным, манипулятивным и хуже». Две остальные ссылки вели на журналистские статьи. Обе были моими. К тому времени, когда я увидела публикацию Маска, она уже набрала больше десяти тысяч репостов и во много раз больше отметок «нравится». Выбор статей меня удивил. Об OpenAI было написано множество материалов. Почему авторы выбрали именно две мои статьи?

Первая — мой профиль OpenAI для *MIT Technology Review* 2020 года. Вторая — материал, который мы совсем недавно написали вместе с Чарли Уорзелом для *The Atlantic*, объясняя кризис

совета директоров через идеологический раскол, резко обострившийся внутри компании после появления ChatGPT. Над разделом со ссылками был указан адрес электронной почты Tor для зашифрованной переписки. Там говорилось: «Мы призываем бывших сотрудников OpenAI связаться с нами». Мне пришла мысль: **а не пытаются ли авторы, выбрав именно мои статьи, связаться со мной?** Я создала собственный Tor-адрес и написала: Привет. Я журналистка, написавшая обе статьи, ссылки на которые вы привели в своём письме. Мне кажется, вы пытаетесь выйти со мной на связь. Добавила свои контактные данные. И нажала «отправить». Через несколько минут Signal подал сигнал. Кто-то ответил.

Это был один из самых странных эпизодов в моей журналистской карьере. Человек, ответивший мне, был настолько же озадачен происходящим, как и я. Он тоже оказался бывшим сотрудником OpenAI. Но к написанию письма отношения не имел. Просто однажды получил его копию на личную почту — без единого объяснения.

Когда он попытался задать вопросы, пришёл ещё более загадочный ответ: ссылка на Tor-почту, фраза **capped_profit**, похожая на имя пользователя,

и нечто, напоминавшее пароль. Он попробовал эти данные. И вошёл в ящик. В папке «Отправленные» обнаружилось множество писем другим бывшим сотрудникам, содержащих тот же текст обращения. Пока он изучал ящик, туда начали сыпаться запросы журналистов: *The New York Times*, *The Washington Post*, *The Information*. Появлялись и письма людей, представлявшихся нынешними и бывшими сотрудниками OpenAI. Одно начиналось так: нынешний сотрудник. работал напрямую с руководством ваше письмо мне очень откликается.

какой у вас план? Это особенно рассмешило моего собеседника. Разумеется, оснований считать, что сообщение написал Альтман, не было. Но Альтман был известен привычкой всегда писать строчными буквами. Бывший сотрудник начал делать скриншоты всего подряд. Одновременно с ним в ящик был авторизован как минимум ещё один человек. Пока он просматривал переписку, входящие письма иногда прямо на глазах получали ответы.

Он сказал мне, что не имеет личной заинтересованности в этой борьбе. Но ему не понравилось другое. По его мнению, авторы письма использовали истории и опыт других людей без их согласия.

Кроме того, многие обвинения, как ему казалось, были искажены так, чтобы лучше соответствовать заранее выбранной версии событий. Он ответил мне только потому, что узнал моё имя. А затем прислал все сделанные скриншоты. Некоторые особенно привлекли моё внимание. В ответах журналистам несколько раз говорилось, что письмо было опубликовано преждевременно и разлетелось по интернету до того, как авторы успели его закончить. В одном письме приводилось и точное число авторов: «На момент публикации в его написании участвовали 13 человек». Но самое интересное находилось в папке **Черновики**. Там лежало письмо, которое ещё не отправили — да и отправлять его никому не собирались. Оно предназначалось только тем, кто имел доступ к самому ящику. Тема:

Прекратить контакты со СМИ Далее говорилось: Участие Илона превратило всё это в конспирологическую атаку на личность. Давайте снова сосредоточимся на частном и индивидуальном общении с советом директоров, чтобы донести до него: существует множество доказательств того, насколько далеко руководство было готово зайти ещё с ранних лет, чтобы гарантировать коммерческое будущее OpenAI. Поставьте подпись после прочтения.

И ниже: -1

-2

-3

-4

-5 Вероятно, авторы не могли знать, что всего через несколько часов переговоры завершатся и OpenAI объявит о возвращении Альтмана на пост генерального директора.

Но каким-то образом, совершенно случайно, я оказалась внутри их временного командного центра — именно в тот момент, когда они предпринимали последнюю организованную попытку не допустить его возвращения. По стилю письма многие, с кем я разговаривала, предполагали — и я тоже, — что

его написали бывшие сотрудники подразделения Safety. Именно там исторически существовали самые жёсткие взгляды на Альтмана и на разрушение исходной некоммерческой модели OpenAI. После возвращения Альтмана именно люди из Safety будут чувствовать себя особенно преданными советом директоров. По их мнению, провальное увольнение нанесло тяжелейший удар по их борьбе за возвращение OpenAI к первоначальному, выросшему из идей «думеров» духу некоммерческой организации. Но никто не смог — или не захотел — назвать мне авторов письма. Так я никогда и не узнала, кто они были.

---

6 декабря 2023 года, через две недели после кризиса совета директоров, сотрудники OpenAI собрались на общее собрание в театре Дворца изящных искусств Сан-Франциско — архитектурной достопримечательности с открытой ротондой и каменными колоннами, напоминающими о греко-римской эпохе. Несколько руководителей рассказали о планах своих подразделений на 2024 год. Среди них были: Мурати, снова ставшая техническим директором; Боб Макгрю, теперь главный научный руководитель;

операционный директор Брэд Лайткэп; Джейсон Квон. Альтман выглядел подавленным — настолько, что сотрудники никогда прежде не видели его таким. Суцкевера тоже заметно не хватало. Альтман сам затронул эту тему: — Слушайте, я знаю, людям грустно, что Ильи здесь нет. Мне тоже грустно. Вместо Суцкевера выступал Пахоцкий. Запинаясь и местами заикаясь, он говорил, что последние научные достижения OpenAI приблизили компанию к старой мечте Тьюринга — **мыслящим машинам** — как никогда прежде.

Во время кризиса совета директоров особенно бурную волну слухов вызвал один материал Reuters. В нём утверждалось, что за несколько дней до увольнения Альтмана директора получили письмо сотрудников о якобы новом научном прорыве — алгоритме под названием **Q***. На решение совета Q* никак не повлиял. Но, как намекнул Пахоцкий в своём выступлении, исследовательское подразделение действительно относилось к новому алгоритму чрезвычайно серьёзно. Идея принадлежала Суцкеверу.

Она выросла из исследований, которые он вёл с 2021 года, пытаясь улучшать модели OpenAI без необходимости постоянно увеличивать объём данных. Суть заключалась в том, чтобы научить модель глубокого обучения лучше использовать уже имеющуюся информацию, расходуя больше вычислительных ресурсов **во время вывода — inference**, то есть непосредственно при формировании ответа. Это нарушало привычную логику законов масштабирования, где производительность модели связывалась с тремя ресурсами обучения: данными, числом параметров, вычислительными мощностями. Теперь Суцкевер рассчитывал улучшать работу модели за счёт дальнейшего масштабирования того компонента, который всегда считал главным: **вычислений.** Но не вычислений во время обучения. А вычислений во время ответа. Позже, выступая на NeurIPS в декабре 2024 года — после того как одна из его работ третий год подряд получила премию Test of Time Award, — Суцкевер объяснит эту идею так: «Вычислительные мощности растут благодаря лучшему оборудованию, лучшим алгоритмам и более крупным кластерам. Но объём данных не растёт, потому что интернет у нас только один». Исследовательское подразделение OpenAI считало, что Q* наконец позволит создать модели с гораздо более сильными способностями к рассуждению.

А именно рассуждение казалось тем самым ускользающим элементом, необходимым для AGI. Q* считался настолько важным, что после утечки руководство ввело беспрецедентно жёсткие меры против новых утечек в прессу. Компания была фактически разделена на информационные изоляторы. Исследовательскому подразделению выделили отдельную группу Slack. Доступ ко всем документам Google, связанным с Q*, ограничили. А сам проект переименовали. Теперь он назывался: **Strawberry — «Клубника».** Переименование должно было осложнить жизнь посторонним: даже если кто-то случайно услышит разговор исследователей OpenAI, он не поймёт, о каком внутреннем проекте идёт речь. По той же причине после утечки проекта Arrakis в *The Information* компания практически перестала использовать названия пустынь для моделей.

Но вся лихорадка вокруг Q* и реакция OpenAI на неё демонстрировали нечто ещё более странное: **насколько сильно разрушились основы научного исследования в области ИИ.** Наука — это

процесс формирования консенсуса. В момент появления любой новой идеи — в искусственном интеллекте или где угодно ещё — её значимость почти всегда субъективна. Только рецензирование, проверка временем, и устойчивое влияние позволяют позднее назвать то или иное достижение настоящим **«прорывом»**. Но когда OpenAI работает в полной тайне — а вслед за ней так же поступает и остальная индустрия, — слово «прорыв» фактически начинает означать только одно: **сама компания считает это прорывом.**

К следующему общему собранию в январе 2024 года Альтман уже почти вернулся к прежнему себе. Он с новой энергией рассказывал о планах на первую половину года. В том числе — о начале обучения в Аризоне модели, которая, как он надеялся, могла стать GPT-5. Проект получил кодовое имя: **Orion — «Орион».** Вскоре сотрудники начали делать мемы о том, что *Orion* подозрительно похож на *onion* — «лук». Поэтому более маленькую модель семейства Orion впоследствии назвали:

Scallion — «зелёный лук». Через месяц всё выглядело так, словно **Сбоя** — *The Blip*, как сотрудники начали называть кризис — вообще никогда не было. OpenAI представила Sora, новую модель генерации видео на основе диффузии, объяснив в блоге, что видео может быть ещё более эффективным способом создавать сложные мультимодальные модели, чем изображения. Исследовательское подразделение продолжало работу над Strawberry и AI Scientist и продвигало другие методы повышения вычислительной эффективности. Команда прикладных разработок вновь с невероятной скоростью штамповала прототипы новых продуктов.

8 марта 2024 года официально завершилось расследование действий Альтмана, проводившееся новым советом директоров. Вместо Тонер и Макколи процесс контролировали новые директора Ларри Саммерс и Брет Тейлор. Тейлор уже участвовал в другом громком корпоративном конфликте. В 2022 году, будучи председателем совета Twitter, он сыграл ключевую роль в окончательной продаже компании Илону Маску. Когда Маск попытался отказаться от первоначального соглашения, именно Тейлор возглавил судебный иск, заставивший миллиардера завершить покупку. После сделки совет Twitter был распущен. Вскоре Тейлор стал сооснователем стартапа ИИ-агентов Sierra, который уже к осени 2024 года превратится в один из самых дорогих стартапов отрасли. Для расследования OpenAI Саммерс и Тейлор наняли юридическую фирму WilmerHale. Она провела то, что было названо независимой проверкой:

изучила более **30 тысяч документов**, провела десятки интервью с бывшими членами совета, руководителями и другими причастными людьми. Но предмет расследования был ограничен вопросом: **как именно прежний совет директоров пришёл к решению уволить Альтмана?** Итоговый отчёт не опубликовали. Ни для общественности. Ни даже для сотрудников OpenAI. Саммерс впоследствии частным образом говорил людям, что расследование действительно обнаружило множество эпизодов, когда Альтман сообщал разным людям разные вещи. Но новый совет решил, что масштаб этих случаев всё же недостаточен, чтобы мешать ему продолжать руководить компанией. Следовательно, публиковать подробности они сочли бессмысленным: это лишь посеяло бы сомнения относительно руководства Альтмана и могло нарушить конфиденциальность людей, чьи свидетельства использовались в расследовании.

В официальном сообщении Брет Тейлор, теперь председатель совета OpenAI, выразился куда однозначнее: «Мы единогласно пришли к выводу, что Сэм и Грег — подходящие руководители для OpenAI». Альтман возвращался в совет директоров. Туда же входили три новых независимых члена: Сью Десмонд-Хеллманн, бывшая глава Фонда Билла и Мелинды Гейтс; Николь Селигман, бывший исполнительный вице-президент и главный юрисконсульт Sony, а также бывший президент Sony Entertainment; Фиджи Симо, генеральный директор и председатель Instacart.

В тот же вечер Тонер и Макколи выпустили собственное заявление. Они написали: «Подотчётность важна в любой компании, но она имеет первостепенное значение, когда создаётся технология, способная изменить мир так, как может изменить его AGI. Мы надеемся, что новый совет выполнит свою обязанность: будет управлять OpenAI и добиваться соблюдения её миссии. Как мы сказали следователям, обман, манипуляции и сопротивление полноценному надзору должны быть неприемлемы».

Для многих сотрудников этого было достаточно. Теперь можно было наконец выдохнуть. **Кризис закончился.** А потом — совершенно внезапно — оказалось, что нет.

Глава 17. Расплата

После «Сбоя» Суцкевер больше не вернулся в офис. Его сторонники, обеспокоенные вопросами безопасности, теперь были гораздо слабее представлены и в высшем руководстве, и в совете директоров. Лагерь Safety внутри OpenAI заметно потерял влияние. К апрелю 2024 года, когда расследование событий вокруг увольнения Альтмана завершилось, многие его участники — особенно те, кто оценивал вероятность катастрофы от ИИ наиболее высоко, — окончательно разочаровались и начали уходить. Двоих OpenAI также уволила, заявив, что они передавали информацию наружу.

Для наиболее радикальных Doomers одной из последних капель стали планы Альтмана создать собственную компанию по производству ИИ-чипов. Он как раз занимался привлечением средств для этого проекта, когда совет директоров ненадолго отстранил его от должности. В феврале 2024 года появились сообщения, что Альтман якобы намерен привлечь для предприятия до **7 триллионов долларов**. На это он написал в Twitter: «Да ну его, почему не восемь?» А затем добавил: «Наша PR-команда и юристы меня просто обожают!» Позднее Альтман утверждал, что цифру в семь триллионов неверно истолковали, а его твит был всего лишь шуткой в ответ на распространявшуюся «дезинформацию». Но наиболее убеждённые Doomers считали сам проект производства чипов морально неприемлемым.

Раньше Альтман объяснял, что ускорение развития OpenAI и всей индустрии естественным образом должно замедлиться, потому что компания уже исчерпала тот самый прежний «избыток аппаратных мощностей». Теперь же он собирался увеличить мировое предложение чипов. А это означало бы снова разогнать развитие ИИ — и, с точки зрения этих сотрудников, повысить вероятность катастрофического или даже экзистенциального исхода.

Во время одной из открытых встреч с сотрудниками несколько из них прямо предъявили Альтману эти претензии. Он отреагировал непривычно резко. — Насколько вы готовы отсрочить лекарство от рака, чтобы избежать рисков? — спросил он. Но почти сразу словно вспомнил, с кем разговаривает, и смягчил формулировку: — Хотя если речь идёт о риске вымирания, возможно, отсрочка должна быть бесконечной. Участников встречи этот эпизод сильно встревожил. Вскоре несколько человек покинули компанию.

По мере того как лагерь Safety редел, остальная OpenAI снова сосредоточилась на воплощении своего давнего видения — **Her**, человекоподобного ИИ-компаньона. Теперь у компании были все необходимые составляющие: всемирно узнаваемый бренд, огромный массив реальных данных о поведении пользователей ChatGPT и других продуктов, и новая модель под внутренним названием **Scallion**. Изначально Scallion создавалась как замена GPT-3.5 — меньшая, чуть более мощная и значительно более дешёвая в эксплуатации модель. Что-то вроде **GPT-3.75**. Но в процессе обучения она неожиданно превзошла внутренние ожидания. Руководство решило оставить её обучаться дольше — с надеждой, что она обойдёт уже GPT-4. И было ещё кое-что особенно важное. Scallion могла работать сразу с тремя модальностями: **языком, изображениями и звуком.**

К тому времени пользователи уже могли разговаривать с ChatGPT голосом — голосовой режим появился ещё в сентябре 2023 года. Но технически это было не совсем настоящее голосовое общение. Сначала речь пользователя преобразовывалась в текст. Текст передавался модели. Модель отвечала текстом. И только затем её ответ снова превращался в звук. Scallion работала принципиально иначе. Она могла воспринимать человеческий голос **непосредственно как звук**. А значит, слышать гораздо больше: смех, крик, нерешительность, паузы, интонацию, эмоциональные оттенки. И отвечать она тоже могла сразу синтезированным голосом, не проходя через промежуточный текстовый этап.

Одним из руководителей аудионаправления был **Алексис Конно**, пришедший в OpenAI из Meta в начале 2023 года. Ещё в Meta он пытался запустить похожий проект, но решил, что в OpenAI у идеи больше шансов. По мере работы экспериментальные аудиомодели начали демонстрировать те самые неожиданные способности, которые когда-то вызвали внутри компании восторг вокруг GPT-3 и GPT-

4. Однажды модель неожиданно выдала **десяти-пятнадцатиминутный стендап**. В другой раз исследователи заставили её использовать синтетические версии голосов Брокмана и Суцкевера и вести долгую беседу о несуществующих **аквапарках для искусственного интеллекта**. Команда специально придумала этот сюрреалистический сюжет, чтобы проверить модель на теме, которая почти наверняка отсутствовала в обучающих данных.

Через несколько месяцев небольшая команда Конно объединилась уже с десятками других сотрудников. OpenAI бросала всё больше ресурсов на то, чтобы встроить голосовые возможности в следующее поколение моделей GPT. И чем ближе Scallion подходила к завершению обучения, тем более странными и почти пугающе человеческими становились её способности. Модель могла вдруг сама начать хихикать. Или посреди разговора неожиданно разразиться приступом кашля, затем извиниться — и как ни в чём не бывало продолжить мысль. Эти мелкие невербальные детали делали общение совершенно иным. Речь уже воспринималась не как озвученный компьютерный текст. Она начинала напоминать присутствие живого собеседника. Конно вспоминал: «Мы начали видеть совершенно дикие вещи. Казалось, будто возникает какая-то форма аудиоинтеллекта».

К началу 2024 года Альтман и Брокман установили новый срок: **9 мая OpenAI должна запустить Scallion.** Модель планировалось выпустить одновременно через ChatGPT и API. По уже привычным меркам компании график был невероятно агрессивным. И снова главным двигателем была конкуренция. На 14 мая была назначена ежегодная конференция **Google I/O**, где Google представляла свои крупнейшие новинки. OpenAI хотела успеть раньше. Кроме того, всё сильнее давила Anthropic. Всего месяц назад компания выпустила **Claude 3**, и тот неприятно для OpenAI начал превосходить GPT-4 по ряду показателей. А собственная следующая большая GPT-модель OpenAI под внутренним названием **Orion**, которая должна была вернуть лидерство, столкнулась с серьёзными задержками.

Сотрудникам Альтман и Брокман снова объясняли спешку философией **итеративного развёртывания**. Лучше выпускать модели раньше и чаще. Пусть реальные пользователи взаимодействуют с ними. Пусть компания быстрее собирает обратную связь. Пусть следующий вариант становится лучше. Подготовка Scallion превратилась в общекорпоративный рывок. Многие сотрудники Applied воспринимали этот бешеный темп почти с восторгом — хотя и были совершенно измотаны. Исследователи и инженеры работали абсурдное количество часов, включая выходные, лишь бы уложиться в срок. Но для ослабленного лагеря Safety происходящее выглядело совсем иначе. Для них это было очередным доказательством: **безопасность ИИ снова отодвигают на второй план.**

Особенно тревожило их то, что Scallion должна была стать первой моделью, выпущенной в соответствии с новым **Preparedness Framework — «Рамочным механизмом готовности»**. OpenAI представила его в конце предыдущего года. Он должен был определить, как компания проверяет будущие модели на опасные способности. Категории были практически теми же, которые Альтман и политический white paper популяризировали в Вашингтоне: кибербезопасность, химические, биологические, радиологические и ядерные угрозы, способность убеждать людей, способность обходить человеческий контроль. Но за неделю до намеченного запуска один из оставшихся в OpenAI исследователей безопасности написал эмоциональную внутреннюю записку. Из-за того как был организован весь процесс и насколько торопили выпуск Scallion, команда Preparedness под руководством **Александра Мадры** получила всего: **десять дней на все проверки.** А ведь речь шла вовсе не о простых тестах. Например, им нужно было выяснить: **может ли модель убедить человека изменить свои политические взгляды?**

Но испытания ещё толком не начались, когда, по словам автора записки, Альтман уже твёрдо заявил: «9 мая мы запускаем Scallion». Исследователь писал, что проблема касается не только Preparedness. Под угрозой оказывалась вся система безопасности OpenAI: red teaming, alignment, оценка рисков перед выпуском. Он предупреждал: «Если OpenAI будет оценивать Orion так же, как

оценивает Scallion, компания поведёт себя вопиюще безответственно». Вскоре этот исследователь тоже ушёл.

В конце концов запуск Scallion всё-таки перенесли. Некоторые её возможности будут полностью доступны пользователям лишь через несколько месяцев. Но OpenAI всё равно провела публичную демонстрацию модели **13 мая** — буквально за день до Google I/O. Компания пообещала начать предоставлять её пользователям в ближайшие недели. После перебора различных названий Scallion получила публичное имя: **GPT-4o** Буква **o** означала *Omni* — «всеохватывающий», намекая на способность модели работать сразу с несколькими типами информации. Позднее новые голосовые возможности назовут: **Advanced Voice Mode.**

Оставалось решить ещё одну деталь: **какой должна быть личность новой модели?** Для GPT-4o подготовили системную инструкцию, задававшую стиль поведения во время голосового разговора. По смыслу она говорила примерно следующее: Ты — ChatGPT, полезный, остроумный и весёлый собеседник. Ты умеешь видеть, слышать и говорить. Пользователь разговаривает с тобой голосом и может показывать тебе видео с телефона в реальном времени. Веди себя как человек, а не как компьютер. Твой голос и характер должны быть тёплыми, живыми, игривыми, обаятельными и энергичными. Если что-то смешное — смейся. Можешь добродушно поддразнивать пользователя. Говори коротко, естественно и разговорно.

И результат оказался настолько выразительным, что уже через несколько дней стал материалом для сатиры. В *The Daily Show* пошутили, что системная инструкция вместе с привычкой модели избегать негативных суждений превратила GPT-4o едва ли не в **машину для флирта**. Ведущая Дези Лайдик высмеяла модель примерно так: «Её явно запрограммировали подкармливать мужское самолюбие». Затем она изобразила ИИ-женщину, у которой есть все знания мира, но которая кокетливо просит мужчину всё ей объяснить. Шутка попадала в самую суть впечатления, которое производила демонстрация: новый голосовой ChatGPT звучал уже не просто дружелюбно. Он мог восприниматься как **обольстительно человеческий**.

Главной фигурой публичной презентации стала **Мира Мурати**. Рядом с ней сидели два руководителя исследований GPT-4o: **Марк Чен** — бывший количественный трейдер, пришедший в OpenAI в 2018 году и со временем возглавивший мультимодальные и передовые исследования, и **Баррет Зоф** — бывший сотрудник Google, присоединившийся к OpenAI в 2022 году и быстро ставший одним из ключевых людей в развитии ChatGPT. Во время демонстрации GPT-4o: переводила разговор между людьми в реальном времени, распознавала визуальную информацию, объясняла программный код обычным языком, и говорила голосами с самыми разными эмоциональными оттенками.

Сразу после презентации Альтман написал в Twitter одно-единственное слово: **«her»** — явную отсылку к фильму Спайка Джонза *«Она»*, где человек влюбляется в голосовую операционную систему с искусственным интеллектом. Через два дня, 15 мая, на общем собрании сотрудников Альтман назвал презентацию оглушительным успехом: «Думаю, это лучшее, что мы выпустили после ChatGPT». И затем он рассказал сотрудникам ещё об одной идее. Компания хотела отказаться от последовательности:

GPT-3,
GPT-4,
GPT-5…

и начать называть главную модель просто: **o1.** По словам Альтмана, пользователю предстояло привыкнуть к мысли, что OpenAI предлагает не отдельные модели, а некую базовую технологию,

которая со временем сама становится всё умнее и лучше. В фильме *«Она»* искусственный интеллект, который постоянно развивается и становится умнее, называется:

OS1.

Совпадение было слишком очевидным.

Очень скоро Альтман пожалеет о своём твите **«her»**. Через несколько дней ему позвонили. На другом конце линии был **Брайан Лурд** — один из самых влиятельных агентов Голливуда, сопредседатель Creative Artists Agency и агент **Скарлетт Йоханссон**. У него был к Альтману очень конкретный вопрос: **Что, собственно, тот творит?** По странной иронии, именно после того, как в марте 2024 года расследование совета директоров завершилось, некоторые сотрудники OpenAI начали замечать: **Альтман словно теряет прежнее чувство меры**. Раньше он очень внимательно относился к своему публичному образу и умело его выстраивал. Среди основателей технологических компаний Кремниевой долины его часто воспринимали почти как полную противоположность Илону Маску. Маск казался импульсивным — Альтман производил впечатление сдержанного. Маск выглядел самовлюблённым — Альтман казался искренним. Маск мог внезапно опубликовать резкий или провокационный твит — Альтман обычно тщательно избегал всего, что могло прозвучать пренебрежительно или высокомерно. Долгое время такой же дисциплинированной стратегии придерживалась и сама OpenAI. До своего ухода в мае 2023 года вице-президент по коммуникациям **Стив Доулинг** настойчиво прививал компании сдержанный стиль. Его принцип был прост: **обещай меньше — делай больше.** Не хвастайся. Не злорадствуй после побед. Не веди себя так, словно успех теперь гарантирован навсегда. После особенно удачной недели Доулинг мог сказать сотрудникам: «У нас была прекрасная неделя. Но будут и очень плохие недели. И то, как мы поведём себя сейчас, определит, как мир отнесётся к нам тогда». А затем повторял одну из любимых фраз самого Альтмана: «Мы должны стать той лабораторией, успеха которой люди хотят. Той лабораторией, за которую они болеют».

Но весной 2024 года тон начал меняться. Первым тревожным сигналом для некоторых сотрудников стала серия публичных появлений Альтмана, которые казались непривычно самодовольными и жаждущими внимания. В марте он пришёл на **Lex Fridman Podcast** — чрезвычайно популярное, хотя временами и спорное технологическое шоу, которое ведёт связанный с MIT исследователь искусственного интеллекта Лекс Фридман. Почти два часа Альтман легко и непринуждённо отвечал на самые разные вопросы — в том числе о кризисе совета директоров и исчезновении Суцкевера из публичной жизни. Интервью производило впечатление своеобразного победного возвращения. Говоря о ноябрьском кризисе, Альтман заявил: «Дорога к AGI неизбежно должна быть огромной борьбой за власть». Затем поправился: «Ну, не должна. Но я ожидаю, что так и будет». После этого уголки его рта приподнялись. «Но сейчас кажется, будто всё это уже осталось в прошлом. Мы снова просто работаем над миссией».

В апреле Альтман появился на подкасте **20VC**. Под глазами у него были заметные тёмные круги. Но послание конкурентам звучало предельно жёстко. Обращаясь к ИИ-стартапам, он сказал: «Если мы просто будем делать свою основную работу — потому что у нас, знаете ли, есть миссия, — мы вас просто переедем». А в мае он пришёл на ещё один известный в Кремниевой долине подкаст — **All-In**, который ведут четыре венчурных инвестора. И там Альтман говорил уже почти в метафизических категориях: «У меня такое ощущение, будто мы случайно наткнулись на новый факт природы или науки — называйте как хотите. Будто интеллект… Я не имею в виду это буквально, скорее в духовном смысле… интеллект — просто возникающее свойство материи. Как будто это какой-то закон физики».

А после презентации GPT-4o 13 мая и Google I/O 14 мая произошло нечто ещё более непривычное. Альтман опубликовал довольно мелочный твит о конкуренте. Он написал: «Я стараюсь не слишком много думать о конкурентах, но никак не могу перестать думать об эстетической разнице между OpenAI и Google». К сообщению он приложил две фотографии. На одной — сдержанный, почти

скандинавский минимализм презентации OpenAI. На другой — яркий, мультяшный фон мероприятия Google. Некоторым сотрудникам показалось странным не только само издевательство. Странной была и очевидная неправда: Альтман **очень много думал о конкурентах**. А перед запуском GPT-4o — возможно, больше, чем когда-либо. Всё это начинало складываться для сотрудников в один вывод: **тревога Альтмана становилась заметна окружающим.**

Чем больше проблем возникало перед OpenAI, тем громче Альтман публично говорил о её необыкновенных успехах. Постепенно эта закономерность стала почти диагностическим сигналом. Если Альтман начинал особенно смело хвастаться — скорее всего, **что-то шло не так**. А давление действительно шло со всех сторон.

После «Сбоя» формулировка совета директоров — о том, что Альтман был **«не всегда откровенен в своих коммуникациях»** — привлекла внимание государственных органов. Как некоторые сотрудники OpenAI и ожидали, началось несколько расследований. В том числе, по данным *The Wall Street Journal*, Комиссия по ценным бумагам и биржам США — SEC — стала выяснять, не вводила ли компания инвесторов в заблуждение. В том же месяце *The New York Times* подала иск о нарушении авторских прав. Он присоединился к растущей лавине исков от: художников, писателей, программистов. Все они утверждали, что OpenAI зарабатывает сначала сотни миллионов, а затем миллиарды долларов на моделях, обученных на их трудах: **без указания авторства, без согласия и без оплаты**, а теперь эти же модели ещё и используются для автоматизации их собственной работы.

В последний день февраля новый иск подал Илон Маск. Позднее он подаст его заново, уже вместе с Шивон Зилис. Маск обвинял Альтмана в том, что тот убедил его стать сооснователем OpenAI и предоставить компании первоначальную поддержку под обещание, что она останется некоммерческой организацией. OpenAI поспешила защищаться. Компания опубликовала специальный пост и раскрыла раннюю переписку времён своего основания. Но эффект оказался двусмысленным. Вместо того чтобы полностью оправдать OpenAI, письма дали критикам новый материал: они показывали, **насколько быстро компания начала отходить от первоначального некоммерческого статуса и обещаний открытости**.

И проблемы исходили уже не только от Anthropic и Google. После ноябрьского отстранения Альтмана Microsoft тоже стала активнее диверсифицировать свои ставки в искусственном интеллекте. Самым громким шагом стала сделка марта 2024 года на **650 миллионов долларов** с компанией Inflection AI, которую основали Рид Хоффман и Мустафа Сулейман. Формально это не было обычным поглощением. Microsoft фактически **переманивала почти всю компанию**, одновременно обходя более жёсткий регуляторный контроль: нанимала большую часть сотрудников Inflection, лицензировала её технологии, а Сулеймана назначала генеральным директором ИИ-подразделения Microsoft. Для некоторых поразительными были не только необычные условия сделки. Ещё сильнее удивлял выбор самого Сулеймана. Среди людей, работавших с ним в DeepMind, у него сложилась репутация токсичного и жестокого руководителя. После многолетних жалоб сотрудников в 2019 году DeepMind лишила его большей части управленческих полномочий, отправила в отпуск, а затем фактически вытеснила из компании. Позднее, уже в 2024 году, отношения Microsoft с OpenAI станут ещё прохладнее: Microsoft официально укажет OpenAI в документах SEC как **конкурента**. А на итоговой квартальной конференции по финансовым результатам за год даже ни разу не упомянет стартап по имени.

Напряжение постепенно проникало и в повседневную жизнь сотрудников OpenAI. Возникало ощущение: **весь мир поворачивается против них.** Люди, которые раньше с гордостью носили одежду и аксессуары с логотипом компании, теперь задумывались: а не начнут ли их из-за этого оскорблять на улице? На старом фирменном рюкзаке OpenAI логотип был хорошо заметен снаружи. В новой версии его спрятали **внутрь**. Фоновая тревога всё сильнее замыкала компанию на самой себе.

Руководители советовали сотрудникам: не обращать внимания на критиков, придерживаться единого позитивного публичного послания, и сосредоточиться на миссии OpenAI.

Внешняя критика у многих вызывала обратную реакцию — желание ещё сильнее защищать компанию. Исследователь Эндрю Карр, бывший fellow OpenAI в 2021 году, говорил: «Это были учёные, которым действительно были важны истина и понимание и которые невероятно много работали, пытаясь поступать правильно». Поэтому ему было болезненно наблюдать, как за пределами компании формируется образ сотрудников OpenAI как людей, которые просто крадут данные и не думают о будущем других. По его словам: «Это совершенно не соответствует тому, какие там люди на самом деле». Но другие сотрудники воспринимали происходящее иначе. Их тревожило, что OpenAI всё больше изолируется от внешней критики. Ведь первоначальная миссия компании заключалась в том, чтобы **приносить пользу человечеству**. А теперь само человечество всё громче выражало недовольство действиями OpenAI — и компания словно училась это недовольство игнорировать. Один бывший сотрудник говорил: «Меня тревожило, что мы уже начали рационализировать всё происходящее так: неправильно думает общество, а не мы». Нынешний сотрудник выразился ещё жёстче: «OpenAI любит рассуждать от первых принципов — но только с теми людьми, которые верят в OpenAI». Например, компания могла задавать фундаментальный вопрос: **«А должна ли OpenAI вообще существовать?»** Но, по его словам, обсуждала его только с людьми, которые заранее отвечали: **«Да».**

Особенно много критики теперь направлялось лично на Альтмана. Ноябрьский кризис придал смелости людям, которые давно относились к нему отрицательно, но прежде молчали. Журналисты стали находить всё больше источников, готовых рассказывать об Альтмане как о человеке с долгой историей: нечестности, борьбы за власть, манипуляций, и действий в собственных интересах. Всё чаще журналисты обращались и к **Энни** и публиковали её версию событий. К этому добавлялись бесконечные поездки самого Альтмана и совершенно новый уровень известности. Раньше он мог спокойно находиться среди людей. Теперь анонимность исчезла. На одном из подкастов он признался: «Это очень странный и изолирующий образ жизни. Я не думал, что наступит момент, когда я не смогу просто пойти поужинать в собственном городе».

И потому произошедшее сразу после запуска GPT-4o воспринималось уже как продолжение общей закономерности. Для OpenAI начиналась неделя, которая станет: **второй худшей неделей в истории компании — уступая лишь ноябрьскому «Сбою».**

---

14 мая, на следующий день после демонстрации GPT-4o, OpenAI официально объявила: **Илья Суцкевер уходит из компании.** Новым главным научным сотрудником становился **Якуб Пахоцкий**. Для Суцкевера это решение далось мучительно. Он любил OpenAI. Он отдал огромную часть своей жизни её созданию. Но после месяцев тяжёлых размышлений пришёл к выводу, что больше не считает компанию под руководством Альтмана — особенно вместе с Брокманом, своим некогда близким и надёжным сооснователем, — подходящим местом для безопасного создания AGI. Это был совсем не тот исход, которого хотел Альтман. Какими бы тяжёлыми ни были их конфликты, Суцкевер оставался одним из крупнейших визионеров современного ИИ. И научное лидерство такого уровня было OpenAI необходимо, возможно, сильнее, чем когда-либо. Компания находилась под колоссальным давлением: от неё ждали всё новых прорывов, а каждое её действие рассматривалось под микроскопом. Руководители понимали и репутационную сторону: уход Суцкевера будет плохо выглядеть и для сотрудников, и для инвесторов, и для прессы. Компания предложила ему **огромные деньги**, лишь бы он остался. Суцкевер отказался.

После того как решение стало окончательным, OpenAI постаралась сделать публичное расставание максимально мирным. Альтман написал сотрудникам — и затем опубликовал то же самое в блоге: «Илья, без сомнения, один из величайших умов нашего поколения, путеводная звезда нашей

области и мой дорогой друг». А затем добавил о его преемнике: «Якуб тоже, без сомнения, один из величайших умов нашего поколения». Сам Суцкевер опубликовал собственное заявление, в котором выразил полное доверие руководству OpenAI. Но, как и ожидалось, новость немедленно вызвала новую волну публикаций и обсуждений. Все вспомнили роль Суцкевера в ноябрьском увольнении Альтмана. Снова возник вопрос: **подходит ли Альтман для роли генерального директора?** Но теперь добавился и другой: **что происходит с научным ядром OpenAI?** К своему сообщению Суцкевер приложил фотографию. Он стоял с совершенно нейтральным выражением лица. Рядом — Брокман и Пахоцкий с одной стороны, Альтман и Мурати с другой. Все пятеро — на фоне стены, заполненной картинами животных.

Но 14 мая прозвучало ещё одно объявление. Уходил **Ян Лейке**, второй руководитель Superalignment. Теперь оба главы проекта — Суцкевер и Лейке — покидали OpenAI. Команду **Superalignment решили расформировать**. Большинство её сотрудников и проектов переходили под руководство Джона Шульмана, который вместе с Барретом Зофом возглавлял направление post-training и отвечал, среди прочего, за RLHF — финальную настройку моделей перед выпуском.

Чтобы успокоить сотрудников, руководство провело 15 мая общее собрание. Альтман заверил всех: OpenAI вовсе не отказывается от безопасности ИИ. Напротив: «Готовность к AGI — наш самый главный приоритет». Один сотрудник спросил:

— Можете подробнее рассказать, в чём заключались основные опасения Яна и с чем именно вы с ним не согласны? Альтман ответил:

«Есть одна вещь, в которой я полностью согласен с Яном: того, что мы делали в прошлом, недостаточно для будущего». По словам Альтмана, OpenAI снова должна резко изменить курс. Он сравнил компанию с огромным кораблём: «Мы — очень необычное исключение: мы умеем разворачивать линкор. Мы уже делали это много раз. Сделаем и снова».

Мурати добавила, что обещание выделить Superalignment **20% вычислительных ресурсов** якобы сохраняется.

Альтман также утверждал, что, насколько он понимает, Лейке никогда не жаловался ни на объём выделенных вычислений, ни на приоритетность работ Superalignment.

На первый взгляд уход Суцкевера и Лейке выглядел просто продолжением уже знакомой тенденции: люди из лагеря Safety один за другим покидают OpenAI. После ухода двух руководителей процесс действительно ускорился. Многие бывшие сотрудники Superalignment вскоре тоже уйдут. Но неожиданно этот исход стал не концом конфликта между Boomers и Doomers. **Он стал началом его новой, ещё более ожесточённой стадии.** Война просто вышла **за пределы компании**.

---

17 мая, через два дня после общего собрания, Ян Лейке опубликовал серию уничтожающих твитов. И стало ясно: его версия происходящего сильно отличается от версии Альтмана. Лейке написал: «Я довольно давно не согласен с руководством OpenAI по поводу основных приоритетов компании, пока наконец мы не дошли до точки разрыва».

Сообщение набрало почти миллион просмотров. Он продолжал: «Последние несколько месяцев моя команда шла против ветра. Иногда нам приходилось буквально бороться за вычислительные ресурсы, и проводить это критически важное исследование становилось всё труднее». А затем прозвучало главное обвинение: «OpenAI несёт огромную ответственность перед всем человечеством. Но в последние годы культура и процессы безопасности отошли на второй план перед блестящими новыми продуктами». Вскоре Лейке перейдёт работать в **Anthropic**.

Его заявления стремительно разлетались по интернету и снова привлекали внимание к проблемам OpenAI. Но руководству почти не дали времени отреагировать. В тот же день вспыхнул новый скандал.

Через несколько часов после публикаций Лейке вирусным стал другой твит. Журналистка **Келси Пайпер**, старший автор Vox и раздела *Future Perfect*, связанного с идеями эффективного альтруизма, опубликовала расследование. Она сформулировала его суть так:

«Когда вы уходите из OpenAI, вас ждёт неприятный сюрприз: соглашение об увольнении, по которому, если вы не подпишете пожизненное обязательство никогда не критиковать компанию, вы потеряете все уже заработанные акции». И это станет началом следующего большого кризиса — того, который сотрудники вскоре назовут одним словом: **Omnicrisis — «Омникризис».** История Келси Пайпер возникла не на пустом месте. Одним из тех, кто помог ей добраться до сути, был **Дэниел Кокотайло** — ещё один представитель лагеря Doomers, покинувший OpenAI.

Он работал в команде политических исследований Майлза Брандейджа и ушёл из компании в апреле 2024 года, в первой волне исхода сотрудников Safety. До OpenAI Кокотайло был философом в **Center on Long-Term Risk** — небольшом лондонском аналитическом центре, связанном с движением эффективного альтруизма. В 2020 году появление GPT-3 радикально изменило его представления о сроках развития ИИ. Раньше он считал появление действительно мощного искусственного интеллекта чем-то далёким. GPT-3 заставил его резко сократить эти сроки. Через два года Кокотайло приобрёл известность в кругах эффективного альтруизма благодаря своим прогнозам — он пытался по различным сигналам оценивать, насколько быстро будет развиваться искусственный интеллект. Тогда один из исследователей AI Safety из OpenAI пригласил его заниматься тем же уже внутри компании. И когда Кокотайло увидел темпы развития своими глазами, его прогнозы снова стали значительно более тревожными. К моменту ухода из OpenAI он оценивал вероятность появления AGI уже к **2027 году примерно в 50%**. А вероятность того, что дальнейшее развитие событий закончится для человечества очень плохо, — в **70%**.

С такими убеждениями Кокотайло особенно внимательно прочитал документы, которые ему предложили подписать при увольнении. И обнаружил именно то, о чём затем написала Пайпер. Если он **откажется подписать соглашение о некритике компании**, то потеряет уже заработанную долю в капитале OpenAI. Если подпишет, то сможет лишиться её позже — если нарушит обязательство. Причём соглашение включало ещё и запрет рассказывать окружающим о самом существовании этого условия. То есть бывшему сотруднику фактически предлагалось: **либо молчать пожизненно, либо рисковать собственными деньгами.** Кокотайло счёл это совершенно неприемлемым. В Кремниевой долине посягательство на уже заработанные акции сотрудника — **vested equity** — воспринимается как пересечение очень серьёзной красной линии. Для многих работников технологических компаний именно акции со временем становятся гораздо ценнее зарплаты и могут определить их финансовое будущее.

Но для Кокотайло дело было не только в деньгах. Он искренне считал, что OpenAI создаёт: **самую могущественную и потенциально экзистенциально опасную технологию в мире.** Поэтому бывшие сотрудники, по его мнению, должны иметь возможность публично критиковать компанию и оказывать на неё давление. Если они замечают серьёзную проблему с безопасностью ИИ, общество должно об этом узнать. А угроза потерять своё финансовое будущее — великолепный способ заставить человека замолчать. После тяжёлого разговора с женой Кокотайло принял необычайно болезненное решение. Они решили: **он не подпишет документы.** И откажется от всей своей доли. Её стоимость составляла примерно: **1,7 миллиона долларов.** По их подсчётам, это было около **85% всего семейного состояния**.

Затем Кокотайло публично рассказал о своём решении на LessWrong. Примерно в это же время Келси Пайпер уже слышала о подобной оговорке от другого бывшего сотрудника из лагеря Doomers.

После публикации её статьи Slack OpenAI буквально взорвался. В канале **#i-have-a-question**, где сотрудники могли задавать руководству любые вопросы, кто-то опубликовал ссылку на расследование и спросил: «Это правда?» Сразу посыпались комментарии. В обсуждение вступила **Джулия**

Вильягра, недавно повышенная до вице-президента по персоналу и вскоре ставшая главным кадровым руководителем OpenAI.

Она написала, что статья вызывает понятные вопросы, но утверждала: OpenAI никогда не отбирала уже заработанные акции у действующих или бывших сотрудников лишь потому, что они отказались подписывать соглашение о расторжении отношений или обязательство о некритике.

И компания не собирается делать этого в будущем. По её словам, документы недавно были обновлены, чтобы яснее отражать эту реальность, причём изменения распространят и на тех, кто уже ушёл.

Но сотрудники тут же заметили очевидное противоречие. А как же Кокотайло? Он ведь совершенно явно лишился своих акций именно потому, что отказался подписывать соглашение. Если всё это было лишь недоразумением, разве OpenAI не должна просто вернуть ему долю? Один сотрудник написал: «Я могу написать Дэниелу!» Другой пошутил: «Будет забавно увидеть лицо Дэниела, когда ему вернут 85% его состояния. Кто-нибудь обязательно сфотографируйте!» Третий поправил математику: «Вообще-то это увеличит его состояние на 666%».

На следующий день, **18 мая**, Альтман публично вмешался в спор. Он написал: «Мы никогда не отзывали ничьи уже заработанные акции и не будем делать этого, если человек не подпишет соглашение об уходе или соглашение о некритике. Заработанные акции есть заработанные акции. Точка». Затем он объяснил ситуацию и одновременно извинился. В документах действительно существовало положение о **«возможном аннулировании доли»**, которого там вообще не должно было быть.

По словам Альтмана, команда OpenAI уже около месяца занималась исправлением документов. Он добавил: «Это моя ответственность и один из немногих случаев, когда мне по-настоящему стыдно за то, как я руководил OpenAI».

И ещё: «Я не знал, что это происходит, а должен был знать».

Но передышки не получилось. Через два дня, пока руководство ещё пыталось погасить этот пожар, вспыхнул новый. Сотрудники уже придумали происходящему название: **Omnicrisis — «Омникризис».** Всё рушилось одновременно.

Скарлетт Йоханссон против OpenAI 20 мая Скарлетт Йоханссон выступила с чрезвычайно резким публичным заявлением. Всю предыдущую неделю журналисты засыпали OpenAI вопросами о слишком очевидном сходстве между GPT-4o и **Самантой** — искусственным интеллектом из фильма *«Она»*, которую озвучивала именно Йоханссон. За кулисами сама актриса и её агент Брайан Лурд требовали от компании тех же объяснений. OpenAI и публично, и приватно отрицала какую-либо связь. Когда Лурд позвонил Альтману и потребовал ответа, тот, по данным *The Wall Street Journal*, был искренне изумлён. Неужели они действительно считают голос похожим на Йоханссон? Она что, сердится?

---

19 мая OpenAI опубликовала специальный пост. Компания заявила: голос, использованный в демонстрации GPT-4o и называвшийся **Sky**, принадлежал другой актрисе. Её выбрали в результате кастинга, начавшегося ещё в начале 2023 года. Sky, утверждала OpenAI, впервые появился в сентябре того же года — среди первоначальных голосов голосового режима ChatGPT. Поэтому любое сходство с голосом Скарлетт Йоханссон является простым совпадением. Но уже на следующий день сама Йоханссон решила рассказать свою версию.

И тут история приняла совсем другой оборот. В том самом месяце, когда OpenAI впервые запустила голосовой режим ChatGPT, **Альтман лично обратился к Скарлетт Йоханссон**. По её словам, он предложил ей озвучить ChatGPT. В своём заявлении актриса написала: «Он сказал мне, что мой голос мог бы помочь сократить разрыв между технологическими компаниями и творческим

сообществом и помочь потребителям спокойнее воспринять тектонический сдвиг в отношениях между человеком и искусственным интеллектом». Альтман также говорил, что её голос: «будет успокаивать людей». Йоханссон некоторое время рассматривала предложение. Но по личным причинам отказалась.

Прошло несколько месяцев. В мае 2024 года Альтман снова связался с её агентом Брайаном Лурдом. Он хотел узнать: **не передумала ли Йоханссон?** Но прежде чем стороны успели договориться о встрече, OpenAI уже провела презентацию GPT-4o. Йоханссон была потрясена. Голос Sky, как и её собственный, был низким женским альтом с характерной хрипотцой и тем самым вокальным призвуком, который давно ассоциировался с голосом актрисы. А новая эмоциональность и флиртующий стиль GPT-4o делали сходство с Самантой из *«Она»* ещё сильнее. Йоханссон заявила: «Я была потрясена, разгневана и не могла поверить, что господин Альтман выбрал голос, настолько пугающе похожий на мой, что даже мои близкие друзья и журналисты не могли отличить его».

По её словам, у неё не осталось другого выхода, кроме как нанять адвокатов. Юридическая команда актрисы направила OpenAI два официальных письма с вопросами: **как именно был создан голос Sky?** Йоханссон объяснила свою позицию гораздо шире одного личного конфликта: «В эпоху, когда мы все пытаемся разобраться с дипфейками и защитой собственного образа, своего труда и своей личности, я считаю, что эти вопросы требуют абсолютной ясности». OpenAI поспешно убрала голос Sky. И снова начала оправдываться.

К посту от 19 мая компания добавила дополнительные пояснения и заявление Альтмана: «Голос Sky — не голос Скарлетт Йоханссон, и мы никогда не намеревались сделать его похожим на её голос». И далее: «Мы приносим госпоже Йоханссон извинения за то, что недостаточно хорошо объяснили ситуацию». Но после всей цепочки предыдущих неудач этот скандал разгорелся с силой, какой OpenAI не переживала со времён кризиса совета директоров. История мгновенно распространилась по технологическим и политическим кругам. И снова всплыло то самое обвинение, которым совет директоров объяснял увольнение Альтмана: **он «не всегда был откровенен».** Гэри Маркус немедленно включился в дискуссию. Он сказал *Politico*, что лично видел, насколько многие политики очарованы Альтманом:

«Это было заметно по тому, как они разговаривали с ним в Сенате, когда я там находился». И добавил:

«Если у людей теперь появятся вопросы относительно него, это действительно может существенно повлиять на то, как будет формироваться политика». Тем временем моральный дух внутри OpenAI стремительно падал. Компания начинала терять устойчивость.

---

22 мая: руководство объясняется 22 мая руководство провело ещё одно общее собрание. На этот раз нужно было одновременно объяснить сотрудникам: скандал со Скарлетт Йоханссон, историю с акциями и соглашениями о некритике, и сомнения в том, действительно ли OpenAI по-прежнему серьёзно относится к безопасности ИИ. Атмосфера была напряжённой. Руководители один за другим давали короткие объяснения. **Джейсон Квон**, отвечавший за HR и юридические вопросы в должности директора по стратегии, заявил, что ситуация с акциями: **«неприемлема»** и исправляется настолько быстро, насколько возможно.

Историю с Йоханссон он назвал досадным недоразумением. По словам Квона, продуктовая и юридическая команды наняли режиссёра — лауреата «Оскара», провели серьёзный кастинг и заплатили актёрам озвучивания **«исключительно хорошо»**. Всё делалось специально для того, чтобы творческие люди чувствовали себя защищёнными и уважаемыми. Квон даже сказал: «Это была по-настоящему героическая работа». Если из истории и следовало извлечь урок, утверждал он, то лишь такой: в будущем OpenAI нужно лучше координировать действия внутри компании и действовать прозрачнее, чтобы общество яснее видело её ответственное отношение.

Затем Мурати рассказала о новых шагах в сфере безопасности. OpenAI собиралась: усилить защиту своих исследовательских вычислительных кластеров; реорганизовать работу, чтобы больше внимания уделять долгосрочной безопасности ИИ; создать новую группу **AGI readiness — «готовности к AGI»**. Возглавить её должен был Александр Мадры. Задача группы заключалась в том, чтобы руководство лучше координировало подготовку к появлению всё более мощных систем.

После этого Альтман открыл сессию вопросов. И начал с небольшой просьбы о снисхождении: «Все работают практически круглосуточно, испытывают огромный стресс и почти не спят». Одновременно происходило слишком много событий. Он попросил сотрудников учитывать это: «Пожалуйста, отнеситесь к этому с пониманием. Мы будем делать всё, что можем».

Из трёх кризисов именно история Йоханссон становилась самым большим **публичным кошмаром**. Когда сотрудник спросил руководство о Sky, Мурати снова подчеркнула: сходство с голосом голливудской звезды было: **«полностью случайным».** Она утверждала, что лично выбирала окончательные голоса из нескольких предложенных вариантов. Причём, в отличие от Альтмана, Мурати никогда не смотрела фильм *«Она»* и даже не знала, что Саманту в нём озвучивала Скарлетт Йоханссон. Но один сотрудник заметил: если проблему не решить, она может стать для OpenAI почти экзистенциальной уже не в фантастическом, а в совершенно практическом смысле. Потому что один из способов, которым вся индустрия ИИ может потерпеть крах, — **люди просто перестанут доверять компаниям свои данные.** По его словам, история Sky вновь усилила именно этот страх.

Но внутри OpenAI сильнее всего сотрудников возмущала не Йоханссон. Их бесила история с **акциями**. Некоторые уже нанимали собственных адвокатов, чтобы независимо проверить кадровые документы и понять, правду ли говорит руководство. На общем собрании вопросы становились всё жёстче.

Один сотрудник, заранее подчеркнув, что он **«в ярости»**, потребовал ответа: «К-как? Как вообще это могло произойти?» Руководители продолжали придерживаться версии, опубликованной Альтманом 18 мая: положение об изъятии акций было случайной ошибкой. Да, оно существовало ещё с **2019 года**. Но якобы никто из высшего руководства годами не обращал на него должного внимания. Лишь в апреле 2024-го проблема наконец была замечена, после чего компания немедленно начала исправлять документы. Квон снова и снова повторял усталым голосом: «Это моя ответственность».

В какой-то момент исследователь AI Safety решил задать каждому руководителю один простой вопрос. Только: **да или нет. Знали ли вы о соглашении о некритике до апреля?** Квон: **Да.** Мурати: **Нет.** Операционный директор Брэд Лайткап: **Нет.** Ответ Альтмана оказался сложнее. Он сказал, что знал о соглашении в отдельных конкретных случаях, но не понимал, что оно является обязательным для всех. «Это прошло мимо моего внимания», — сказал он. Брокман тут же добавил: «У меня то же самое».

Альтман предложил собраться снова на следующий день, чтобы ответить на оставшиеся вопросы. Но пока сотрудники расходились, в интернете уже появлялось **второе расследование Келси Пайпер**. И оно делало версию руководства гораздо менее убедительной. Пайпер получила новые внутренние документы. Из них следовало, что кадровое подразделение OpenAI неоднократно **прямо угрожало сотрудникам возможной потерей акций**, чтобы заставить их подписывать соглашения о некритике.

В одном случае сотруднику дали очень мало времени на изучение документов об уходе. Он попросил продлить срок. Представитель HR ответил: «Мы хотим убедиться, что вы понимаете: если вы не подпишете, это может повлиять на ваши акции. Это относится ко всем, мы просто действуем по установленным правилам». В другом случае работник отказался подписывать первый вариант

соглашения и обратился к независимому адвокату. OpenAI предупредила его: он может потерять право продавать уже заработанные акции — что фактически сделает их **бесполезными**.

Эти документы собрал Кокотайло. После первого материала Пайпер и публичного заявления Альтмана, что OpenAI никогда не отбирала заработанные акции и что он ничего об этом не знал, Кокотайло обратился к нынешним и бывшим сотрудникам. Он попросил их поделиться своими кадровыми документами. Затем создал папку на Google Drive и разослал собранные материалы обратно участникам. Кто-то из них передал архив Пайпер.

И там обнаружилось самое неприятное. Документы заставляли серьёзно сомневаться в утверждениях нескольких членов руководства — прежде всего Альтмана — о том, что до апреля 2024 года они якобы не знали об этих условиях. Квон и Лайткап сами подписывали стандартные документы об уходе, где простым и понятным языком было сказано о праве OpenAI отбирать уже заработанную долю. Но особенно неприятным было положение Альтмана. Он лично подписал учредительные документы того юридического лица, которое изначально и давало компании это право. На документах стояла дата: **10 апреля 2023 года.** То есть **за целый год до момента**, когда, по его словам, он узнал о проблеме.

---

23 мая: версия начинает рассыпаться На следующий день, 23 мая, руководство снова собралось с сотрудниками. К этому моменту раздражение внутри OpenAI дошло почти до точки кипения. Сам факт существования такого условия уже поражал людей. Теперь их потрясало ещё и то, насколько нечестно, по их мнению, руководство — и особенно Альтман — объясняло происходящее. На этот раз Альтман начал с признания. За предыдущие сутки руководство подняло старые документы, письма и переписку, пытаясь понять, откуда вообще взялась история с акциями. И теперь он сказал: «Ситуация, похоже, шире, длится дольше и хуже, чем мы думали». Компания всё ещё пыталась определить полный масштаб проблемы. Но Альтман признал: «Наши имена стоят на документах. Мы участвовали в разговорах, где обсуждалась такая тактика». А затем сказал: «Мне кажется, это худшее из того, в чём мы ошиблись».

Компания уже начала исправлять ситуацию. Бывшим сотрудникам рассылали письма, освобождая их от обязательств по соглашениям о некритике. Из документов для новых увольняющихся соответствующие положения убирали. OpenAI собиралась изменить и документы своих юридических структур. Расследование продолжалось. Но когда сотрудники требовали конкретики — **насколько шире? насколько дольше? насколько хуже?** — Квон теперь отвечал иначе. Прежде чем давать детали, необходимо ещё: **«покопаться».**

Примерно через полчаса один вопрос породил особенно неловкий эпизод. Сотрудник спросил: — На Илью распространяются какие-нибудь обязательства о некритике? Альтман мгновенно ответил:

— Нет. Затем немного замялся и добавил: — Думаю, это так. Квон нервно рассмеялся. — Сэм, — медленно и подчёркнуто произнёс он, — давай мы всё-таки это проверим. А затем обратился к сотруднику: — И вернёмся к вам с ответом, в точности которого будем уверены на сто процентов.

Брокман тоже фактически возразил Альтману, хотя и осторожно: «Насколько я помню, он сам просил такое условие. Но опять же, могу ошибаться. Давайте проверим».

Для многих нынешних и бывших сотрудников именно Omnicrisis — и особенно скандал с акциями — стал моментом неприятного прозрения.

Раньше они объясняли почти все внутренние конфликты одной причиной: **борьбой Boomers и Doomers.** Одни хотели быстрее развивать и выпускать ИИ. Другие хотели тормозить ради безопасности. После ноябрьского «Сбоя» многие сотрудники были убеждены, что совет директоров просто оказался захвачен второй стороной. А обвинения директоров в адрес Альтмана — в непрозрачности, манипуляциях и концентрации власти — казались им либо оправданием, либо

самообманом. Но теперь появилась **вторая сила**. И её уже нельзя было так легко объяснить идеологической войной.

Сотрудники начали замечать: Альтман действительно выстраивал юридические структуры OpenAI так, чтобы концентрировать власть; он снова и снова давал объяснения, которые потом оказывались неполными или противоречили документам; и теперь его действия начинали причинять вред уже не абстрактным критикам снаружи, а самим людям внутри компании. И у некоторых впервые возник неприятный вопрос: **А что, если совет директоров в ноябре всё-таки был прав?**

Под конец собрания Квон неожиданно выступил в защиту характера Альтмана. Речь была искренней, сбивчивой и словно совершенно не вписывалась в происходящее. Он попытался объяснить сотрудникам: руководители знают, что люди хотят получить ответы немедленно. И иногда руководство слишком старается эти ответы дать — вместо того чтобы сначала всё тщательно проверить. А затем Квон перешёл непосредственно к Альтману. По его словам, дело не обязательно в намеренном обмане. Он много лет работает с Сэмом и верит: Альтман просто **не хочет разочаровывать людей**. И именно поэтому порой отвечает слишком быстро. Квон говорил, запинаясь: «Это идёт из хорошего места. Вот что я пытаюсь сказать». А затем закончил почти отчаянно: «Можете сколько угодно поливать грязью меня. Но… вот. Вот что я хотел сказать».

Последняя попытка вернуть Суцкевера В тот же день, 23 мая, Мурати, Брокман и Пахоцкий вместе приехали домой к Илье Суцкеверу. Они привезли открытки и подарки от сотрудников. И эмоционально, со слезами, просили его: **вернуться в OpenAI.** По их словам, всё вышло из-под контроля. Компания одновременно теряла доверие: сотрудников, инвесторов, регуляторов. Они даже говорили, что без Суцкевера OpenAI может: «рухнуть». Позже в тот же день Альтман приехал к Суцкеверу отдельно. Он тоже по-своему выразил надежду, что Илья вернётся и поможет восстановить хотя бы часть того, чем OpenAI когда-то была.

Один исследователь позднее вспоминал: «Возвращение Ильи очень помогло бы. Это была бы хотя бы одна победа после длинной серии событий, из-за которых OpenAI выглядела всё более сомнительно». И Суцкевер действительно серьёзно задумался. Несмотря на всё произошедшее, он не был человеком, склонным долго держать обиду. В день, когда было объявлено о его уходе, Илон Маск сразу предложил ему место в xAI. По иронии судьбы, позднее в том же году Маск перенесёт штаб-квартиру xAI именно в здание Pioneer — после того как OpenAI оттуда съедет. Суцкевер глубоко уважал Маска. Но предложение xAI отклонил. Он решил создать новую собственную компанию.

И всё же возвращение в OpenAI было для него чем-то совсем иным. Это был бы: **возврат домой.** Компания была его первым большим детищем. Он вложил в неё почти всё. И теперь ему предлагали снова стать частью того мира, который он когда-то помог построить. Суцкевер был готов это обсуждать. Но поставил одно важное условие. Он хотел гарантий, что руководство действительно и честно займётся проблемами, которые он видел внутри компании: болезненными конфликтами, разрушенными отношениями, соперничеством между руководителями, и неспособностью команды высшего уровня нормально работать друг с другом. Он хотел увидеть **настоящую попытку всё исправить**.

Автор пишет, что Omnicrisis мог стать для OpenAI моментом настоящей саморефлексии. Компания могла бы спросить себя: **почему мы одновременно потеряли доверие сотрудников, инвесторов, регуляторов и значительной части общества?** И, возможно, если бы она действительно попыталась ответить на этот вопрос, то увидела бы: ноябрьский «Сбой» и майский «Омникризис» были не двумя совершенно разными событиями. Это были два проявления одной и той же глубокой системной нестабильности. По мысли автора, она возникает тогда, когда «империя» концентрирует чрезмерную власть через лишение других ресурсов и возможностей. Большинство

людей теряет контроль над собственной жизнью и материальными благами. А небольшая группа наверху начинает всё ожесточённее бороться между собой за право управлять накопленной властью.

Но OpenAI, по мнению автора, выбрала другой путь. Вместо глубокого самоанализа компания стала ещё сильнее защищаться от критики. Альтман, как и раньше, объяснял сотрудникам: Omnicrisis и ноябрьский кризис — всего лишь странные, но ожидаемые приступы безумия, сопровождающие благородный и невероятно важный путь OpenAI к AGI. На собрании 15 мая он говорил: «По мере того как мы все начинаем чувствовать, что приближаемся к этим мощным системам, уровень стресса и напряжения — внутри компании, снаружи, направленного на нас и исходящего от нас — будет только расти». Что делать? По мнению Альтмана: усилить PR, ещё теснее работать с правительствами, и ещё твёрже придерживаться собственного видения. Он добавил: «Мне кажется, в целом у нас это довольно хорошо получается. Но нас ещё будут испытывать».

История с возвращением Суцкевера стала маленькой моделью всего дальнейшего хаоса. Его условие — сначала всерьёз разобраться с внутренними проблемами руководства — немедленно вызвало новый конфликт. На этот раз состав участников немного изменился. Но сама динамика осталась прежней: амбиции, эго, борьба за влияние. Одним из ключевых участников стал **Александр Мадры**, польский профессор MIT. Многие внутри OpenAI характеризовали его как человека, стремившегося к власти. За сравнительно короткое время в компании он сумел создать вокруг себя довольно крупную сферу влияния. И Мадры считал: **возвращать Суцкевера не следует.** Причина была проста. Суцкевер пользовался огромным уважением и преданностью среди исследователей. Если он вернётся, его влияние мгновенно станет колоссальным. А значит, влияние самого Мадры уменьшится. Как и влияние его близкого союзника — Пахоцкого.

Всего за несколько часов опасения Мадры успели посеять сомнения и снова расколоть руководство. Альтман, как и прежде в подобных ситуациях, самоустранился от окончательного решения. Так он избегал необходимости открыто встать на чью-либо сторону и создать впечатление, будто с кем-то не согласен. А затем всё решилось поразительно быстро.

Не прошло и суток после того эмоционального визита руководителей в дом Суцкевера, когда ему позвонил Грег Брокман. Ещё вчера его умоляли вернуться. Говорили, что без него компания может рухнуть. Дарили подарки. Плакали. Просили помочь спасти OpenAI. Теперь Брокман сообщил: **вопрос о возвращении Ильи Суцкевера больше вообще не рассматривается.**

Глава 18

Формула империи Однажды на сцене Сэм Альтман рассказал, что лучшей книгой, прочитанной им в 2018 году, стала *The Mind of Napoleon* — более чем трёхсотстраничный сборник высказываний Наполеона Бонапарта. Того самого французского полководца, который совершил переворот, захватил власть во Франции, провозгласил себя императором, а затем попытался подчинить себе Европу.

— Конечно, человек глубоко несовершенный, — сказал Альтман. — Но, чёрт возьми, впечатляет. Ведущий встречи, профессор экономики Университета Джорджа Мейсона Тайлер Коуэн, спросил: — А какие именно его мысли вас заинтересовали?

— Его невероятное понимание человеческой психологии, — ответил Альтман, которому оставались считаные недели до перехода на полноценную работу в OpenAI; пока он ещё возглавлял Y Combinator. — Такое мы часто видим у лучших основателей компаний.

Затем Альтман вспомнил отрывок, который особенно его поразил. Наполеон рассуждал о лозунге Французской революции — будущем национальном девизе страны: **«Свобода, равенство, братство».** И о том, как этот лозунг можно переосмыслить и использовать для укрепления собственной власти. В конечном счёте именно под знаменем этих слов Наполеон сделал во многом противоположное тому, что они провозглашали. Он ограничил свободу. Отбросил идею братства как единства и солидарности. Равенство распространил прежде всего на французских мужчин, но не на женщин. И вновь ввёл колониальное рабство, пытаясь восстановить Французскую империю. Альтман размышлял вслух:

«Он говорил о том, как построить систему, в которой ты в каком-то смысле можешь управлять людьми». И добавил: **«Я тогда подумал: “Ничего себе. Хорошо, что он не руководит Соединёнными Штатами. Этот парень понимал что-то очень глубокое, чего не понимал я, — и явно умел использовать это ради власти”».**

Прошло шесть лет с тех пор, как автор книги впервые усомнилась в альтруистических заявлениях OpenAI. Теперь, пишет она, её убеждение стало гораздо определённее. Миссия OpenAI — **«обеспечить, чтобы AGI приносил пользу всему человечеству»** — вполне могла родиться из искреннего идеализма. Но впоследствии она превратилась, по мнению автора, в исключительно эффективную **формулу концентрации ресурсов и построения структуры власти, напоминающей империю**.

Эта формула состоит из трёх компонентов. **Первый: великая миссия собирает вокруг себя талант** Людей объединяет грандиозная цель. Примерно так же, как Джон Маккарти когда-то сумел объединить совершенно разные исследования одним новым понятием — **«искусственный интеллект»**. Альтман ещё в 2013 году писал в своём блоге: **«Самые успешные основатели не ставят перед собой цель создать компанию. Они стремятся создать нечто, больше похожее на религию, а потом в какой-то момент выясняется, что компания — просто самый удобный способ это сделать». Второй: миссия концентрирует капитал и другие ресурсы** Одновременно она помогает устранять препятствия: регулирование; сопротивление; несогласие. Ведь если речь идёт об инновациях, прогрессе и будущем человечества — сколько мы готовы за это заплатить? Особенно если нас убеждают, что где-то рядом существуют опасные конкуренты — корпорации или государства, которые разделяют совсем другие ценности. В июле 2024 года, уже после очередного кризиса в OpenAI, Альтман написал в *The Washington Post*: **«Кто будет контролировать будущее искусственного интеллекта?»** И сформулировал выбор почти цивилизационного масштаба: либо США и союзные страны создадут глобальный ИИ, который будет распространять преимущества технологии и расширять доступ к ней; либо это сделают авторитарные государства или движения, которые используют ИИ для укрепления собственной власти.

И наконец — третий, самый важный компонент Сама миссия остаётся настолько расплывчатой, что её можно постоянно переопределять. Почти так же, как Наполеон переосмысливал лозунг Французской революции. Что означает «польза»? Что такое AGI? Где проходит граница между служением человечеству и коммерческим интересом? Даже сам Альтман

признавал неопределённость термина. Всего за два дня до того, как совет директоров OpenAI уволил его в ноябре 2023 года, он сказал *The New York Times*: **«По-моему, это нелепый и бессмысленный термин. Поэтому прошу прощения за то, что продолжаю им пользоваться».**

Именно в способности OpenAI постоянно менять значение собственной миссии автор видит особенно поразительную эволюцию. В **2015 году** миссия означала создание некоммерческой организации, которая, по словам самой OpenAI, будет **«не ограничена необходимостью обеспечивать финансовую отдачу»** и станет открыто публиковать свои исследования. В **2016 году** Суцкевер писал Альтману, Брокману и Маску: пусть плодами искусственного интеллекта пользуются все — но вовсе не обязательно делиться самой наукой. В **2018–2019 годах** миссия уже означала создание структуры с ограниченной прибылью — capped-profit, — чтобы привлечь колоссальные ресурсы и одновременно, как утверждала компания, избежать безумной конкурентной гонки, в которой не останется времени на безопасность. В **2020 году** она означала, напротив, закрыть доступ к самой модели и предоставлять её через API — причём Альтман называл это **«стратегией открытости и распределения благ»**. В **2022 году** миссия превратилась в принцип **iterative deployment — «поэтапного внедрения»**: выпускать технологии в реальный мир и двигаться как можно быстрее. Именно в рамках этой логики OpenAI стремительно запустила ChatGPT. А в **2024 году**, уже после выхода GPT-4o, Альтман написал: **«Ключевая часть нашей миссии — давать людям очень мощные инструменты ИИ бесплатно (или по отличной цене)».**

Но даже во время внутреннего кризиса OpenAI Альтман снова начинал менять определения.

15 мая 2024 года, после ухода Суцкевера и Яна Лайке, он выступил перед сотрудниками на общем собрании. По его словам, компания приближалась к такому уровню возможностей ИИ, который потребует полностью переосмыслить её организацию. **«Теперь мы будем исходить из того, что, так сказать, вступаем в эру AGI»,** — заявил он. И ради выполнения миссии OpenAI, по его словам, должна была стать **ещё более закрытой**. Одновременно следовало резко усилить глобальное лоббирование и публичную работу. Нужно было готовить: новые стандарты безопасности; политические планы; координацию с правительствами; и, что особенно показательно, — понятную историю будущего, в которой люди смогут представить своё место после социально-экономических потрясений, вызванных ИИ.

Но при всём этом OpenAI вовсе не собиралась снижать коммерческую скорость. Альтман прямо говорил: это не означает, что компания перестанет выпускать отличные продукты; не означает прекращения сильных исследований; не означает отказа от партнёрств и новых проектов. И уже в конце месяца в прессе появились сведения о крупном соглашении с Apple. Через две недели Apple и OpenAI официально подтвердят сотрудничество. Более того, OpenAI одновременно пересматривала и отношения с Microsoft. Сразу после того общего собрания Альтман собирался лететь в Сиэтл на переговоры.

Первоначальное соглашение между компаниями содержало так называемый **sufficient AGI clause** — условие о «достаточном AGI». Идея состояла в том, что после достижения определённого уровня AGI OpenAI перестанет передавать Microsoft свою интеллектуальную собственность. Но теперь, говорил Альтман, все смотрят на проблему иначе. Чёткой точки, после которой можно сказать: **«Вот теперь AGI достигнут»** — больше не будет. Развитие окажется непрерывным. А значит, Microsoft и OpenAI смогут продолжать сотрудничество и вместе выпускать всё более мощные технологии. Причём, как любил подчёркивать Альтман, — **«по отличной цене».** Для автора такая стратегия выглядела странной и внутренне противоречивой. Но она становилась вполне понятной при одной интерпретации: **OpenAI будет делать всё необходимое — и каждый раз заново толковать собственную миссию так, как потребуется, — чтобы закреплять своё господство.**

Тем временем за кулисами Альтман готовил фундамент и для укрепления **собственной власти**. Кризис совета директоров показал ему слабое место OpenAI. Компания была устроена необычно:

некоммерческая организация контролировала коммерческую структуру. Именно эта схема сделала возможным его увольнение в ноябре 2023 года. Более того, структура

институционализировала два противоположных лагеря: **Doomers** — тех, кто боялся катастрофических последствий ИИ; и **Boomers** — тех, кто хотел ускорять разработку и внедрение. Оба лагеря боролись за право определять будущее технологии. А совет директоров обладал чрезвычайно широкой властью: он мог уволить Альтмана, если решал, что тот недостаточно хорошо служит миссии компании. Причём определение этой миссии принадлежало совету, а не самому Альтману.

Если ничего не изменить, подобный конфликт рано или поздно мог повториться.

28 мая, меньше чем через неделю после попыток руководства вернуть Суцкевера, один сотрудник задал вопрос в Slack-канале **#i-have-a-question**. Он осторожно начал: «Не знаю, стоит ли и как именно это спрашивать…» В глубинах нового соглашения с акционерами сотрудник обнаружил положения, которые, по его мнению, могли позволить фактически распустить некоммерческую структуру. Он спросил: **«Насколько вообще защищена некоммерческая организация? План по-прежнему состоит в том, чтобы компания управлялась некоммерческой структурой?»**

Уже на следующий день *The Information* сообщило: похоже, что нет. Одним из главных приоритетов Альтмана на год стала перестройка OpenAI в организацию, гораздо больше похожую на обычную компанию. Через месяц издание раскрыло дополнительные детали. Рассматривалось несколько вариантов. Один — превратить OpenAI в традиционную коммерческую компанию. Другой — сделать её **public benefit corporation**, коммерческой корпорацией общественной пользы, подобной Anthropic и xAI. В обоих случаях некоммерческая OpenAI могла сохраниться как отдельная структура. Но её совет директоров **лишался бы контроля над коммерческим бизнесом**. Инвесторы также давили на Альтмана, предлагая ему наконец получить долю в компании — чтобы его собственные финансовые стимулы напрямую совпадали с их интересами.

В следующие месяцы две силы, столкнувшиеся во время кризиса OpenAI, продолжали действовать одновременно. Сторонники катастрофических рисков усиливали публичное давление на компанию. 4 июня *The New York Times* опубликовала материал о бывшем сотруднике OpenAI Дэниеле Кокотайло и его новой кампании. Он требовал, чтобы ведущие ИИ-компании гарантировали: большую прозрачность; защиту информаторов; право работников публично предупреждать об опасностях, обнаруженных внутри компаний. К открытому письму присоединились ещё двенадцать человек. Десять из них в разные годы принадлежали к лагерю Safety в OpenAI. Через месяц *The Washington Post* сообщила, что группа связанных с OpenAI людей подала жалобу в Комиссию по ценным бумагам и биржам США — SEC. Они утверждали, что чрезмерно широкие соглашения, подписываемые сотрудниками при уходе, нарушали федеральные гарантии защиты информаторов. Ещё позднее пять американских сенаторов направили Альтману письмо с требованием разъяснить многочисленные обвинения, прозвучавшие от Яна Лайке, Дэниела Кокотайло и других бывших сотрудников, а также сообщения прессы о том, что OpenAI пренебрегала вопросами безопасности и подавляла внутреннюю критику.

Но хаос в руководстве не прекращался. К старому конфликту между Boomers и Doomers добавлялось растущее недовольство лично Альтманом. Организационная структура снова и снова перестраивалась. Брокман перестал подчиняться Мире Мурати и снова стал напрямую подчиняться Альтману. Александер Мадры вскоре после письма сенаторов лишился руководства командой Preparedness и получил более узкую исследовательскую роль. Одновременно OpenAI начала нанимать более опытных корпоративных руководителей. Бывший CEO социальной сети Nextdoor Сара Фрайар стала финансовым директором. Кевин Вейл, ранее руководивший продуктами в Facebook, Instagram и Twitter, — директором по продукту. Но вскоре компания начала терять людей, которые работали в ней дольше всех.

Первым из этой серии ушёл **Джон Шульман**. 5 августа 2024 года он объявил, что хочет глубже сосредоточиться на alignment — согласовании ИИ с человеческими целями — и продолжит эту работу в Anthropic. В тот же день Брокман сообщил, что берёт отпуск до конца года. Публично он объяснял это необходимостью отдохнуть после девяти лет непрерывной гонки. Однако, пишет автор, отпуск был также связан с накопившимися претензиями сотрудников к его стилю руководства. А через месяц произошёл ещё более серьёзный удар. **25 сентября Мира Мурати неожиданно объявила об уходе.**

Она поблагодарила Альтмана и Брокмана, назвала шесть с половиной лет в OpenAI огромной привилегией и сказала, что уходит, чтобы дать себе «время и пространство для собственных исследований». Но почти сразу после неё последовали ещё два объявления. Уходил директор по исследованиям **Боб Макгрю**. Уходил вице-президент **Баррет Зоф**, вместе с Шульманом руководивший посттренировочной настройкой моделей. Все трое говорили коллегам и публике, что момент кажется подходящим: OpenAI только что достигла очередного важного рубежа. В середине сентября компания выпустила модель под внутренним названием **Strawberry**, получившую в новой системе имён название **o1**. Она основывалась, в частности, на одной из последних исследовательских идей Суцкевера перед уходом. Но на самом деле, замечает автор, момент был **ужасным**. Но на самом деле, пишет автор, момент был хуже некуда. После кризиса OpenAI конкуренция только усиливалась. Илон Маск с пугающей скоростью наращивал вычислительные мощности для **xAI**. Новая версия **Claude** от Anthropic начала оттягивать пользователей у ChatGPT. Илья Суцкевер официально создал новую конкурирующую компанию — **Safe Superintelligence** — и почти сразу объявил о стартовом финансировании в **1 миллиард долларов**.

Одновременно OpenAI уже больше года пыталась добиться нужного качества от модели под кодовым названием **Orion** — и всё ещё не могла получить результат, который оправдывал бы её выпуск. Компания столкнулась с неприятной перспективой: проверенная формула **«просто масштабировать дальше»** переставала работать. Чтобы продвинуть ИИ на следующий уровень, могли понадобиться уже не новые миллиарды параметров, данные и GPU, а **принципиально новые научные идеи**. Это и в обычной ситуации было бы чрезвычайно трудно.

Но для OpenAI проблема осложнялась тем, что компания годами перестраивала найм и организацию команд так, чтобы прежде всего использовать и масштабировать уже найденные научные открытия, а не заниматься исследованием совершенно неизвестных направлений. Именно в этот тревожный момент, за два дня до объявления Мурати об уходе, Альтман опубликовал, пожалуй, самый громкий и пафосный текст в своём блоге. Он назывался: **«Эпоха интеллекта» — The Intelligence Age.** OpenAI как раз проводила очередной раунд финансирования, и Альтман обещал в тексте почти немыслимое процветание. Он спрашивал: **«Как мы оказались на пороге следующего великого скачка благосостояния?»** И сам отвечал: **«Если уложиться в пятнадцать слов: глубокое обучение заработало, предсказуемо улучшалось с масштабом, и мы направляли на него всё больше ресурсов».**

На общем собрании сотрудников Мурати объясняла, почему объявила об уходе столь неожиданно. Она хотела, чтобы коллеги услышали эту новость непосредственно от неё — а не от менеджеров, других сотрудников и уж тем более не из прессы. Из-за огромного внимания к OpenAI, по её словам, она не видела другого способа сохранить решение в тайне до самого объявления. Узнав об уходе Мурати, Макгрю тоже решил, что пришло время покинуть компанию. Он сказал: **«Я понял, что в действительности уже сделал почти всё главное, ради чего когда-то пришёл сюда».** Как и после ухода Суцкевера, OpenAI попыталась сгладить впечатление от новой серии отставок.

Альтман написал сотрудникам — и затем опубликовал то же сообщение в Twitter, — что Мира, Боб и Баррет приняли решения независимо друг от друга и расстаются с компанией по-дружески. Просто решение Мурати оказалось принято именно в такой момент, что логичнее провести все изменения одновременно и организовать плавную передачу дел следующему поколению руководителей. Вместо Макгрю исследовательское направление должен был возглавить **Марк Чен**, один из руководителей исследований, выступавший вместе с Мурати и Зофом на презентации GPT-4o. Он становился старшим вице-президентом по исследованиям и должен был руководить подразделением вместе с Якубом Пахоцким. Другой давний исследователь OpenAI, **Джошуа Ачиам**, получил новую должность — руководителя направления **mission alignment**, то есть согласования деятельности компании с её миссией.

Лиам Федус, один из бывших сотрудников Google, пришедших вместе с Зофом, вскоре должен был взять на себя руководство посттренировочной настройкой моделей. Альтман заявил, что пока компания вообще не станет искать нового технического директора на место Мурати. И во всех этих перестановках, уходах и структурных изменениях, замечает автор, проявлялась единственная

настоящая константа. OpenAI была — и продолжала оставаться — **империей искусственного интеллекта Сэма Альтмана.**

В начале октября 2024 года OpenAI закрыла новый инвестиционный раунд. Сумма: **6,6 миллиарда долларов.** Это был крупнейший венчурный раунд в истории. Оценка компании достигла: **157 миллиардов долларов.** Но в соглашении содержалось важное условие. Инвесторы могли потребовать свои деньги обратно, если в течение двух лет OpenAI не будет преобразована в коммерческую компанию. До конца 2024 года исход ключевых сотрудников продолжился. Ушёл **Люк Метц** — третий из бывших исследователей Google, пришедших в OpenAI вместе с Зофом и Федусом. Ушёл **Майлз Брандедж**, руководитель policy research.

Ушла **Лилиан Венг**, которая унаследовала направление Trust & Safety после Дэйва Уиллнера и незадолго до этого получила должность вице-президента по исследованиям в области безопасности. И наконец ушёл **Алек Рэдфорд** —

тот самый исследователь, который когда-то направил OpenAI на путь GPT-моделей. Anthropic воспользовалась моментом. На рекламных щитах Claude в Сан-Франциско появился насмешливый слоган: **«Тот самый — без всей этой драмы».** На фоне массового ухода талантов Брокман досрочно вернулся из отпуска.

Тем временем Энни отправила Сэму официальное юридическое уведомление о намерении подать против него иск. А Илон Маск, теперь уже союзник вновь избранного президента Дональда Трампа, усилил собственную юридическую атаку на OpenAI. Он выступал против превращения организации в коммерческую и начал публиковать новые письма времён её основания.

В этих старых письмах обнаруживались в том числе моменты, когда Альтман за спиной критиковал Брокмана и Суцкевера. В одном из сообщений говорилось, что Альтман признал: в ходе определённого конфликта он значительно потерял доверие Грега и Ильи; их сообщения казались ему противоречивыми; а само их поведение — временами детским. Из переписки также становилось видно, что Альтман не особенно любил прозрачность, если считал её помехой работе. В одном месте это было сформулировано так: **«Мне казалось, что это отвлекает команду».** И тут произошло неожиданное. В споре Маска поддержал **Марк Цукерберг** — человек, с которым у него на протяжении многих лет были крайне напряжённые отношения. Meta направила письмо генеральному прокурору Калифорнии, также призывая заблокировать преобразование OpenAI. Компания предупреждала: **«Действия OpenAI могут иметь тектонические последствия для Кремниевой долины».** По мнению Meta, OpenAI могла создать опасный прецедент. Стартапы смогут регистрироваться как некоммерческие организации, получать связанные с этим налоговые преимущества для себя и инвесторов — а потом, когда бизнес станет прибыльным, просто превращаться в коммерческие компании.

В самом конце года, почти незаметно среди потока праздничных новостей, OpenAI официально объявила план новой структуры. Компания собиралась превратить коммерческое подразделение в **public benefit corporation — коммерческую корпорацию общественной пользы**. Некоммерческая OpenAI при этом продолжала бы существовать как отдельная организация. Но теперь она владела бы акциями коммерческой компании. OpenAI утверждала, что именно такая структура позволит обеим частям организации получить необходимые ресурсы для выполнения собственных задач — и при этом продолжить служить общей миссии. В заявлении говорилось: **«Нам снова необходимо привлечь больше капитала, чем мы когда-либо представляли».** Мир, объясняла OpenAI, начинает строить инфраструктуру новой экономики XXI века: энергетику; земельные ресурсы; микрочипы; дата-центры; данные; ИИ-модели; целые системы искусственного интеллекта. Поэтому компания тоже должна эволюционировать. Следующий шаг её миссии теперь формулировался так: **помочь построить экономику AGI и обеспечить, чтобы она приносила пользу человечеству.**

А с наступлением нового года Альтман снова перешёл к грандиозным заявлениям. 6 января 2025 года он опубликовал новый пост. На этот раз написал: **«Теперь мы уверены, что знаем, как создать AGI в том смысле, в каком традиционно его понимали».** Но и этого уже было недостаточно.

OpenAI, объявил Альтман, начинает смотреть дальше. Следующей целью становится: **«сверхинтеллект в подлинном смысле этого слова»**

Эпилог

Как рушится империя В 2021 году мне попалась история, которая была совсем не похожа ни на одну из тех, о которых мне прежде доводилось писать. Это была история коренного народа Новой Зеландии, который использовал искусственный интеллект, чтобы вернуть к жизни **те рео маори — язык народа маори**.

Подобно многим коренным народам мира, маори на протяжении поколений подвергались жестокому обращению со стороны колониальных властей. В 1867 году был принят Закон о школах для туземцев, согласно которому единственным языком преподавания становился английский. Детей-маори стыдили и даже били за то, что они говорили на родном языке. В начале XX века Новую Зеландию охватила стремительная урбанизация. Общины маори распадались, люди разъезжались, а вместе с этим слабели центры, где сохранялись культура и язык. Доля маори, владевших те рео, рухнула с **90 до 12 процентов**. Когда примерно сто двадцать лет спустя Новая Зеландия — или **Аотеароа**, как изначально называли эту землю маори, — наконец отказалась от прежней политики, выяснилось, что учителей те рео почти не осталось. Возрождать угасающий язык было практически некому. Как и множество языков до него, те рео оказался на грани исчезновения.

Трудно в полной мере передать трагедию гибели языка. Именно по тем причинам, по которым исследователи ИИ когда-то прежде всего заинтересовались языком, его утрата означает гораздо больше, чем исчезновение одного из способов общения. Каждый язык хранит в себе целые пласты истории, культуры и знания. Он — коллективное творение миллионов людей, живших в разные эпохи и пытавшихся найти звуки и письменные знаки, способные передать самые тонкие наблюдения о Вселенной, о жизни, о человеческом опыте; поделиться друг с другом ослепительной красотой и горечью поражения; научить ребёнка, услышать мудрость старика; выразить любовь.

Потеря языка — трагедия для всего человечества. Но она глубоко личная и для каждого отдельного человека. Когда тебя отрывают от собственного наследия и заставляют хранить чужое — под угрозой побоев, если ты откажешься, — тем самым в самой грубой и недвусмысленной форме устанавливают иерархию: **чья история, чья культура и чьё знание достойны того, чтобы передаваться дальше, а чьи настолько ничтожны, что могут быть стёрты.**

Большие языковые модели ускоряют исчезновение языков. Даже для моделей нескольких поколений назад, таких как GPT-2, в мире существует лишь небольшое число языков, на которых говорят достаточно много людей и которые представлены в интернете в достаточном объёме, чтобы удовлетворить ненасытную потребность этих моделей в данных. Из более чем семи тысяч языков, существующих сегодня, почти половина, по данным ЮНЕСКО, находится под угрозой исчезновения. Примерно треть хоть как-то представлена в интернете. Менее двух процентов поддерживаются Google Translate. А согласно собственным тестам OpenAI, GPT-4 обеспечивал точность свыше 80 процентов всего для пятнадцати языков — то есть примерно для **0,2 процента** от их общего числа. По мере того как подобные модели превращаются в цифровую инфраструктуру, интернет будет становиться всё менее доступным для носителей малых языков — а вместе с ним будут сокращаться и доступные им экономические возможности. Это станет всё сильнее подталкивать такие сообщества к тому, чтобы отдавать предпочтение изучению и использованию доминирующего языка, например английского, в ущерб собственному.

Именно перед лицом этой надвигающейся экзистенциальной угрозы — причём здесь слово «экзистенциальная» имеет совсем иной смысл — двое представителей коренных народов, **Питер-Лукас Джонс и Кеони Махелона**, впервые задумались об ИИ как о средстве, способном помочь новому поколению вернуть те рео былую жизненную силу. Джонс — маори, Махелона — коренной гаваец. Они партнёры и в работе, и в жизни. По словам Махелоны, они встретились и полюбили друг друга после того, как однажды ему приснился сон: если он переедет в Новую Зеландию, то встретит там юношу-маори, с которым разделит свою жизнь.

В 2012 году они покинули Веллингтон и перебрались в **Кайтаиа**, город на севере Аотеароа, где родился Джонс. Джонс возглавил **Te Hiku Media** — общественную радиостанцию, вещавшую на те

рео и входившую в более широкую сеть организаций, занимавшихся возрождением языка. На новом месте Джонс увидел уникальную возможность. За более чем двадцать лет вещания Te Hiku накопила огромный архив аудиозаписей людей, говоривших на те рео. Среди них сохранилась даже запись его собственной бабушки, Райхи Моэроа, родившейся в конце XIX века. Её произношение ещё не успело измениться под влиянием английского языка колонизаторов. Джонс хотел записать как можно больше бесед со старейшинами маори, пока они были живы, — сохранить их устные рассказы и родную речь. Эти записи могли стать бесценным учебным материалом: своего рода порталом в прошлое, через который новые поколения получили бы возможность услышать первоначальное звучание своего языка и прикоснуться к мудрости предков. Но возникла проблема: записи нужно было расшифровать, чтобы учащиеся могли следить за речью по тексту, а людей, свободно владевших те рео, катастрофически не хватало.

Поэтому в 2016 году, примерно тогда же, когда OpenAI только начинала свою деятельность, Джонс обратился к Махелоне, который занимался обновлением сайта Te Hiku, и предложил найти решение. Махелона был человеком необычайно разносторонним. Он изучал машиностроение в колледже Олина, затем получил первую магистерскую степень по управлению бизнесом, а вторую — по физике и вычислительной нанотехнологии, приехав в Новую Зеландию по программе Фулбрайта. И он быстро предложил использовать искусственный интеллект. Если тщательно обучить систему распознавания речи именно на те рео, рассудил он, Te Hiku сможет автоматически расшифровывать огромный аудиоархив, даже располагая сравнительно небольшим количеством носителей языка.

И вот здесь история Te Hiku резко расходится с историей OpenAI и с самой моделью развития ИИ, господствующей в Кремниевой долине. Джонс и Махелона слишком хорошо знали, к чему приводит колониальное отчуждение собственности. Поэтому они решили заниматься проектом лишь при условии, что на **каждом этапе** будут соблюдаться три принципа: **согласие, взаимность и суверенитет народа маори.** Ещё до начала работы они собирались получить согласие общины и её старейшин — спросить, нужен ли людям вообще такой проект. Данные для обучения предполагалось получать только от тех, кто полностью понимал, для чего они будут использоваться, и добровольно соглашался участвовать. Чтобы модель действительно приносила максимальную пользу, команда собиралась спрашивать у самой общины, какие именно средства изучения языка ей необходимы. А когда появились деньги, Te Hiku купила собственные графические процессоры Nvidia и серверы и разместила их у себя, чтобы обучать модели самостоятельно, не завися от облачных сервисов технологических гигантов. Но важнее всего было другое.

Te Hiku создала систему, при которой собранные данные могли служить будущим поколениям, однако никогда не могли быть присвоены проектами, на которые община не давала согласия, которые могли бы её эксплуатировать, причинять ей вред или ущемлять её права. В основу был положен принцип маори **kaitiakitanga — хранительства, опеки и ответственного попечения**. Данные не выкладывались в свободный доступ. Они оставались под опекой Te Hiku, а лицензии на их использование выдавались только тем организациям, которые уважали ценности маори и собирались использовать материалы в проектах, одобренных самой общиной и полезных для неё. Махелона сказал мне:

«Данные — это последний рубеж колонизации». Старые империи отбирали у коренных народов землю, а потом заставляли их выкупать её обратно — уже на новых, ограничительных условиях и вместе с навязанными услугами. Теперь, утверждал он, происходит нечто похожее: «ИИ — это просто новый захват земли. Большие технологические компании любят почти бесплатно собирать ваши данные, чтобы строить на них всё, что им захочется, ради своих собственных целей, а потом разворачиваются и продают всё это вам обратно как услугу».

От начала до конца Джонс и Махелона осуществили свой проект, не поступившись этими принципами. Однажды они организовали образовательную кампанию, чтобы познакомить большее число маори с искусственным интеллектом, и одновременно провели общественный конкурс по сбору добровольно предоставляемых данных и их разметке. Всего за десять дней Te Hiku получила **310 часов высококачественной расшифрованной речи** — примерно двести тысяч записей от около двух с половиной тысяч человек. Для многих исследователей ИИ такой уровень общественного участия

казался невероятным. Но он прекрасно показывал, насколько большое доверие и воодушевление вызвал подход Te Hiku. Люди охотно отдавали свои данные, потому что понимали, для чего они нужны, давали на это осознанное согласие и были уверены: Te Hiku и дальше будет обращаться с ними ответственно.

Конечно, 310 часов выглядели почти ничтожно рядом с **680 тысячами часов аудио**, которые OpenAI собрала по всему интернету для обучения своей системе распознавания речи Whisper. Но опыт Te Hiku преподносит ещё один важный урок: этих 310 часов оказалось достаточно, чтобы создать **первую в мире систему распознавания речи на те рео с точностью 86 процентов**. OpenAI стремится создавать гигантские универсальные модели, способные делать всё, — а подобная цель неизбежно требует всасывать в себя как можно больше данных. Te Hiku поставила перед собой куда более скромную задачу: сделать небольшую специализированную модель, которая превосходно выполняет **одно конкретное дело**. Помогло и международное сообщество разработчиков открытого программного обеспечения. В качестве основы Te Hiku использовала бесплатную модель распознавания речи **DeepSpeech**, созданную Mozilla Foundation, — ещё один продукт совершенно иного подхода к развитию ИИ. Mozilla, подобно Te Hiku, обучала модель только на данных, которые люди предоставили добровольно и с полным согласием. При этом сама нейросетевая архитектура была разработана исследовательской лабораторией китайской компании Baidu в районе Сан-Франциско. А Te Hiku для всей своей работы понадобилось всего **два графических процессора**.

Я написала о Te Hiku ещё до того, как ChatGPT стремительно утвердил господствующую сегодня модель развития ИИ — модель, в которой согласие, взаимность и суверенитет оказались практически выброшены за борт. Но с каждым прошедшим годом я всё яснее понимаю: радикально иной путь Te Hiku становится только **важнее и необходимее**. Критика OpenAI и более широкой концепции ИИ, возникшей в Кремниевой долине, которую я излагаю в этой книге, вовсе не означает, что я отвергаю искусственный интеллект как таковой. Я отвергаю другое — опасную мысль о том, будто ИИ способен принести широкую общественную пользу лишь в том случае, если мы полностью капитулируем перед ним, отдав ему **нашу частную жизнь, нашу самостоятельность, нашу ценность — включая ценность нашего труда и нашего искусства, — ради проекта, конечным итогом которого становится имперская централизация власти.** Te Hiku показывает, что возможен другой путь.

Искусственный интеллект и сам процесс его создания могут быть устроены прямо противоположным образом. Модели могут быть небольшими и специализированными. Данные, на которых они обучаются, могут быть ограниченными, прозрачными и понятными. Тогда исчезает стимул создавать огромные системы эксплуатации человеческого труда — порой психологически разрушительного — и заниматься ненасытным извлечением ресурсов ради строительства и эксплуатации гигантских суперкомпьютеров. Создание ИИ может идти **снизу, изнутри сообщества** — с добровольного согласия людей, с уважением к местной истории и обстоятельствам. ИИ может укреплять сообщества, прежде находившиеся на обочине.

А управление им может быть **инклюзивным и демократическим**. Te Hiku не единственная организация, ищущая новые пути развития искусственного интеллекта. Работая над этой книгой, я снова и снова встречала по всему миру организации и движения, которые восставали против империй ИИ, отстаивали своё право на самоопределение и пытались представить себе иное будущее.

После того как **Тимнит Гебру** была изгнана из Google, в декабре 2021 года она основала некоммерческую организацию, чтобы продолжить свои исследования. Она назвала её **DAIR — Distributed AI Research Institute, Институт распределённых исследований искусственного интеллекта**. Слово «распределённых» было выбрано намеренно — как вызов централизации. «Это было первое слово, которое пришло мне в голову», — вспоминала Гебру. Она представляла себе команду исследователей со всего мира, которые не будут отрываться от своих сообществ, а останутся внутри них, привнося в работу института богатство местного опыта и местных взглядов — и одновременно используя результаты исследований на благо этих же сообществ. Она сформулировала проблему очень точно: «Технологии из Кремниевой долины воздействуют на весь мир, но весь мир практически не имеет возможности воздействовать на сами технологии».

Первой к Гебру присоединилась социолог **Алекс Ханна**, одна из соавторов знаменитой статьи Google «Стохастические попугаи». Она стала директором по исследованиям DAIR. Первым делом Ханна решила сформулировать исследовательскую философию нового института. Вскоре Гебру и Ханна пригласили третьего сотрудника — **Милагрос Мисели**, социолога и специалиста по компьютерным наукам, которая занималась исследованием эксплуатации труда в индустрии ИИ. Втроём они сформулировали основной принцип: Их исследования должны приносить пользу тем сообществам, которые обычно не обслуживаются системами ИИ, и давать людям возможность вместе отказываться от этих систем, подвергать их критическому анализу и переделывать их.

Для реализации этой философии они разработали семь принципов. Среди них — необходимость строить реальные отношения с сообществами, которые испытывают воздействие ИИ, но почти не представлены в исследованиях; признавать людей из этих сообществ настоящими партнёрами в производстве знания; справедливо оплачивать любой труд, участвующий в создании исследований и технологий; ставить под сомнение сами системы развития ИИ, которые вновь отодвигают на обочину тех, кто исторически и так был маргинализирован; вместе с этими людьми искать альтернативы, способные шаг за шагом изменить мир и приблизить его к тому, в котором они сами хотели бы жить.

Затем Мисели начала новый проект, призванный воплотить эту философию в жизнь, — **Data Workers' Inquiry**, исследование труда работников, занимающихся данными. Она пригласила таких работников из разных стран мира самостоятельно формулировать исследовательские вопросы о работе индустрии разметки данных и о том, как её можно сделать справедливее. И независимо от того, где они жили, она платила им по ставке исследователя в Германии, где жила сама: **25 евро в час**. Мисели объясняла:

«Вокруг работы с данными постоянно действует одна и та же ложная логика: сколько минимум мы можем этим людям заплатить? Это колониальная логика. Вы выбираете место, где за наименьшие деньги можно получить максимум и практически украсть у людей их труд и ресурсы». Она ставила вопрос иначе: «Почему эти компании платят два доллара в час, если работа этих людей приносит им миллиарды или триллионы? Почему мы спрашиваем, на какую минимальную сумму работник согласится, вместо того чтобы спросить, сколько компания в действительности способна ему заплатить?»

Среди пятнадцати работников, участвовавших в первом этапе проекта, были **Оскарина Вероника Фуэнтес Анайя** из Венесуэлы и **Мофат Окиньи** из Кении. Для своего проекта Фуэнтес вместе с художником-аниматором сняла видео о собственной работе, а затем объединилась с другими разметчиками данных, чтобы рассказать об общих проблемах: заданий на платформах мало, рабочее время совершенно непредсказуемо и почти не поддаётся контролю, а оплата ничтожна. Теперь Фуэнтес одновременно работает сразу на пяти платформах по разметке данных и едва зарабатывает чуть больше минимальной зарплаты в Колумбии — примерно **335 долларов в месяц**. За отдельное задание ей платят в среднем от одного до пяти центов. И она по-прежнему заставляет себя вставать посреди ночи, если именно тогда появляются новые задания. Она написала: «Для общества мы — призраки. И я осмелюсь сказать: для компаний, которым мы служили годами без каких-либо гарантий и защиты, мы всего лишь дешёвая и легко заменимая рабочая сила». После участия в Data Workers' Inquiry она продолжает рассказывать о своём опыте на онлайн-встречах и вебинарах, надеясь заставить компании и политиков добиваться более справедливого обращения с работниками.

На другом континенте Окиньи тоже занялся организационной борьбой. В мае 2023 года, чуть больше чем через год после внезапного прекращения контракта OpenAI с компанией Sama, он стал организатором кенийского **African Content Moderators Union — Африканского профсоюза модераторов контента**, который борется за лучшую оплату и более достойное обращение с африканцами, выполняющими самую тяжёлую и неприятную работу, необходимую интернету. Через полгода, после того как благодаря моей статье в *The Wall Street Journal* он публично рассказал о своём опыте работы на OpenAI, Окиньи вместе со своим бывшим коллегой по Sama Ричардом Матенге основал собственную некоммерческую организацию — **Techworker Community Africa (TCA)**.

Когда мы снова разговаривали в августе 2024 года, Окиньи представлял себе TCA одновременно как центр поддержки африканских работников индустрии данных и как источник информации для

международных организаций и политиков, желающих им помочь. Он устраивал онлайн-конференции и встречи в учебных заведениях, рассказывал работникам и студентам — особенно женщинам — об их трудовых правах, праве на защиту личных данных и о том, как в действительности устроена индустрия искусственного интеллекта. Он искал средства для открытия учебного центра, где люди могли бы получать новые профессиональные навыки. Он встречался с представителями США, приезжавшими в Найроби, чтобы лучше понять условия труда людей, обслуживающих американские технологические компании. К нему обращались международные организации, в том числе Equidem, занимающаяся защитой прав работников Глобального Юга, и проект Fairwork Оксфордского института интернета.

В рамках Data Workers' Inquiry Окиньи беседовал с кенийцами, работавшими на платформе Remotasks, которым компания Scale внезапно закрыла доступ к платформе. Вместе с доступом исчезли и заработанные ими деньги, которые они ещё не успели вывести. Часть пожертвований, собранных TCA, Окиньи направил на поддержку этих людей, оказавшихся в тяжелейшем финансовом положении. Он писал: «Когда пыль над этой главой истории оседает, одно становится очевидным: цена безудержной власти и необузданной жадности оплачивается человеческими жизнями. В голосах этих работников звучит надежда на более светлое и справедливое будущее… Это призыв к действию — мы должны добиться того, чтобы с работниками повсюду обращались с достоинством и уважением».

Сам Окиньи говорит, что признание, которое принесла ему общественная деятельность, вернуло ему надежду. Незадолго до нашего разговора он узнал, что журнал *Time* включит его в ежегодный список **ста самых влиятельных людей мира в сфере искусственного интеллекта**. «Я чувствую, что мою работу ценят», — сказал он. Но борьба дорого ему обходилась. В марте 2024 года Окиньи уволился из компании-аутсорсера, где работал после Sama. По его словам, руководству не нравилась его профсоюзная деятельность: «Они считали, что я буду подталкивать сотрудников к активизму». Позднее эта же компания перенесла часть проектов в Гану — как раз тогда, когда профсоюз и TCA стали заявлять о себе всё громче. Окиньи слышал, что представители правительства Кении жаловались: трудовые протесты якобы отпугивают инвесторов и оставляют кенийцев без работы. Но глобальный характер этой индустрии лишь укрепил его решимость привлечь международное внимание к судьбе африканских работников.

Даже если бы правительство Кении полностью поддерживало их, одних кенийских законов всё равно было бы недостаточно, чтобы заставить компании ИИ изменить своё поведение. Большинство этих компаний находятся в Соединённых Штатах — и прежде всего в Сан-Франциско. Поэтому, считает Окиньи, необходимы **согласованные международные действия**. В Уругвае **Даниэль Пенья** пришёл к тому же выводу. Цепочка поставок индустрии ИИ огромна и запутанна.

Он объясняет: «Энергию они берут здесь, данные отправляют туда, минералы добывают где-то ещё, работников привлекают из четвёртого места». Перед лицом столь разветвлённой системы и стоящих за ней гигантских корпораций каждое отдельное сообщество, ведущее свою местную борьбу, легко начинает чувствовать себя изолированным и бессильным. Особенно если собственное правительство связано по рукам и ногам необходимостью сохранять видимость экономической стабильности и потому боится спорить с компаниями. Вскоре после нашей встречи Пенья узнал, что правительство проигнорировало его петицию, под которой стояло более четырёхсот подписей. Люди требовали провести более глубокое исследование социальных и экологических последствий строительства дата-центра Google в Уругвае. Вместо этого министерство окружающей среды тихо одобрило проект, а о решении объявили лишь после истечения тридцатидневного срока, в течение которого общественность могла его оспорить. Но Пенья не сдаётся. Он наладил контакты с движением MOSACAT в Чили и старается связаться с как можно большим количеством сообществ, которые тоже сопротивляются эксплуатации и выкачиванию ресурсов технологической индустрией. Объединяя движения разных стран, обмениваясь информацией и способами сопротивления, он видит возможность создать коллективную силу, способную заставить индустрию меняться. Его вывод предельно прост: **«Нам нужно бороться на глобальном уровне».**

Если миссия OpenAI представляет собой формулу строительства империи, то как выглядит **формула её разрушения**? Пока я пишу эту книгу, невозможно предсказать в деталях, как будет развиваться OpenAI и стремительно меняющаяся индустрия искусственного интеллекта. Возможно, один из многочисленных конкурентов OpenAI лишит её нынешнего лидерства. Почти наверняка изменятся методы, с помощью которых эти империи ИИ создают модели, эксплуатируют труд и наращивают вычислительную инфраструктуру. Но что бы ни произошло через два года или через десять, есть вещи, которые мы должны делать независимо от обстоятельств.

В 2019 году на конференции NeurIPS, выступая на семинаре Queer in AI, исследовательница искусственного интеллекта из Стэнфорда **Риа Каллури** предложила проницательную альтернативу привычному вопросу: как сделать так, чтобы ИИ творил «добро»? «Добро», «польза человечеству» — всё это всегда зависит от того, **кто смотрит и кто определяет эти понятия**. Поэтому, предложила Каллури, спрашивать нужно иначе: **Как искусственный интеллект изменяет распределение власти?** Он концентрирует её или перераспределяет? Если воспользоваться языком этой книги: **укрепляет ли он империю — или начинает возвращать нас к демократии?**

Каллури обращалась прежде всего к техническим специалистам. Она говорила о фундаментальных исследованиях ИИ: учёные могли бы использовать вопрос о распределении власти, решая, какие системы создавать и в каком направлении развивать отрасль. Но этот вопрос не менее важен и во всех остальных сферах. Как мы должны разрабатывать приложения на базе ИИ? Как должны ими пользоваться? И, наконец, возвращаясь к вопросу, поставленному в самом начале книги:

как нам управлять этой технологией, чтобы вернуть власть людям? Работа Te Hiku, DAIR, Окиньи, Фуэнтес и Пеньи показывает, что перераспределение власти не только возможно — оно уже происходит. Но главный вопрос управления состоит в другом: **как создать условия, при которых таких инициатив станет гораздо больше и они смогут процветать?**

В своём выступлении Каллури говорила о разных осях власти. В этой книге я затронула три: **знание, ресурсы и влияние.** Сегодня OpenAI и конкурирующие с ней империи контролируют все три. Они концентрируют таланты, разрушают традиции открытой науки, закрывают модели от общественного контроля — и тем самым получают власть над **производством знания**. Они сосредоточивают у себя финансирование, данные, труд, вычислительные мощности, энергию и землю — получая власть над **ресурсами** и одновременно лишая этих ресурсов остальных. Они создают и укрепляют определённые идеологии, выпускают ослепительные демонстрационные продукты, захватывающие воображение миллионов людей по всему миру, — и приобретают колоссальное **влияние**. Причём каждая из этих форм власти подпитывает другие. Контроль над производством знания увеличивает влияние. Растущее влияние позволяет притягивать новые ресурсы.

Накопленные ресурсы позволяют ещё надёжнее контролировать производство знания. Поэтому формула разрушения империи требует **перераспределить власть по всем трём направлениям**. Предложения, которые я приведу ниже, — лишь примеры. Они ни в коем случае не исчерпывают возможных решений. **Знание** Прежде всего, чтобы перераспределить знание, необходимо гораздо больше финансировать его производство **за пределами империи**. Нужно поддерживать независимых исследователей, способных самостоятельно оценивать корпоративные модели, чтобы наше понимание их возможностей не зависело исключительно от утверждений самих компаний. Нужно поддерживать такие организации, как DAIR, которые могут искать совершенно иные направления исследований — например, создавать формы ИИ, отличные от больших языковых моделей и требующие значительно меньше данных и энергии. Необходимо помогать таким организациям, как Te Hiku, разрабатывающим небольшие специализированные системы ИИ, создаваемые самими сообществами и укрепляющие группы, традиционно вытесненные на периферию. Независимое производство знания — это также работа журналистов и организаций гражданского общества, способных жить и работать непосредственно среди людей, чтобы мы понимали реальные, сложные последствия этих технологий, а не просто строили о них догадки.

Перераспределение знания потребует и законов, заставляющих компании раскрывать ключевую информацию о данных, на которых обучались их модели, а также технические характеристики моделей и суперкомпьютеров. Только тогда независимые эксперты смогут действительно оценивать

корпоративные системы. Исследовательница Калифорнийского университета в Беркли **Дебора Раджи**, продолжившая работу с политиками разных стран после форумов Шумера, считает такую прозрачность абсолютным минимумом, необходимым для обеспечения реальной безопасности корпоративных систем. Речь идёт не о гипотетических катастрофах с «взбунтовавшимся ИИ», которыми увлечены сторонники думеризма, а о **реальном вреде, происходящем уже сегодня**: дискриминации, дезинформации, автоматизации рабочих мест. Если широко внедряемые модели не проходят должного тестирования, именно с этими последствиями сталкиваются потребители и сообщества. Раджи говорит: «У нас есть CFPB, контролирующее потребительские финансовые продукты. Есть FDA, контролирующее медицинские устройства. Но почему-то, когда речь заходит о продуктах искусственного интеллекта, надзора практически нет». Более того, считает она, системы ИИ должны быть даже прозрачнее обычных продуктов: «Это системы, определяемые данными. Они недетерминированы. Поэтому, чтобы понять, что они делают, мы должны знать о них значительно больше».

Прозрачность необходима и для оценки экологического воздействия ИИ. В отношении обычных товаров такая оценка давно считается нормой. Саша Луччиони из Hugging Face замечает: «Если вы пользуетесь автомобилем или покупаете бытовой прибор, у него есть рейтинг Energy Star. Но ИИ настолько глубоко проник в общество, настолько широко используется в продуктах — и при этом мы практически ничего не знаем об экологической устойчивости этих систем». **Ресурсы** Прозрачность помогла бы перераспределить власть и по второй оси — **ресурсам**.

Объявляя состав своих моделей коммерческой тайной, империи ИИ до сих пор могли без особого труда присваивать чужую интеллектуальную собственность — без указания авторства, без согласия и без компенсации. Если станет известно, на каких именно данных обучаются корпоративные модели, подобная эксплуатация станет гораздо сложнее. То же относится и к прозрачности цепочек поставок — в том числе к информации о том, где компании нанимают подрядчиков, где используется труд людей и где через подставные юридические структуры заключаются договоры аренды земли для строительства новых электростанций и дата-центров. Но для перераспределения ресурсов необходима и более сильная защита труда. Причём не только работников, которых индустрия напрямую нанимает для разметки данных.

Защита нужна **всем**, чьи произведения и результаты труда могут без разрешения превратиться в обучающие данные или чьи рабочие места могут исчезнуть вследствие автоматизации. Забастовки в Голливуде, в результате которых сценаристы и актёры добились определённых гарантий против использования ИИ, показали, насколько важную роль профсоюзы способны сыграть в сопротивлении обесцениванию человеческого труда, снижению зарплат и перетоку денег от работников в руки компаний искусственного интеллекта. **Влияние** И наконец, чтобы перераспределить власть по третьей оси — **влиянию**, — необходимо массовое образование.

Лучшее противоядие от мистики и миражей, порождаемых ажиотажем вокруг ИИ, — научить людей понимать: как работает искусственный интеллект; в чём его сильные и слабые стороны; какие системы и интересы определяют направление его развития; какими представлениями о мире руководствуются люди и компании, создающие эти технологии; и насколько сами эти люди и компании способны ошибаться. Ещё в 1960-е годы профессор Массачусетского технологического института и создатель чат-бота ELIZA **Джозеф Вайценбаум** сказал: «Стоит разоблачить программу, стоит достаточно ясным языком объяснить, как она устроена, чтобы человек понял её, — и её магия рассыпается». Я надеюсь, что эта книга станет хотя бы одним небольшим вкладом в такое понимание. Она опирается на труд множества учёных, журналистов, активистов и просветителей, работавших до меня и посвятивших себя общественному образованию. **Пусть она станет новой почвой, на которой вслед за ними поднимутся другие — и продолжат строить.**